

# Knowledge Transfer for Melanoma Screening with Deep Learning

Afonso Menegola<sup>†‡</sup>, Michel Fornaciali<sup>†‡</sup>, Ramon Pires<sup>°</sup>,  
Flávia Vasques Bittencourt<sup>•</sup>, Sandra Avila<sup>†</sup>, Eduardo Valle<sup>†\*</sup>

<sup>†</sup>RECOD Lab, DCA, FEEC, University of Campinas (Unicamp), Brazil

<sup>°</sup>RECOD Lab, IC, University of Campinas (Unicamp), Brazil

<sup>•</sup>School of Medicine, Federal University of Minas Gerais (UFMG), Brazil

## ABSTRACT

Knowledge transfer impacts the performance of deep learning — the state of the art for image classification tasks, including automated melanoma screening. Deep learning’s greed for large amounts of training data poses a challenge for medical tasks, which we can alleviate by recycling knowledge from models trained on different tasks, in a scheme called *transfer learning*. Although much of the best art on automated melanoma screening employs some form of transfer learning, a systematic evaluation was missing. Here we investigate the presence of transfer, from which task the transfer is sourced, and the application of fine tuning (i.e., retraining of the deep learning model after transfer). We also test the impact of picking deeper (and more expensive) models. Our results favor deeper models, pre-trained over ImageNet, with fine-tuning, reaching an AUC of 80.7% and 84.5% for the two skin-lesion datasets evaluated.

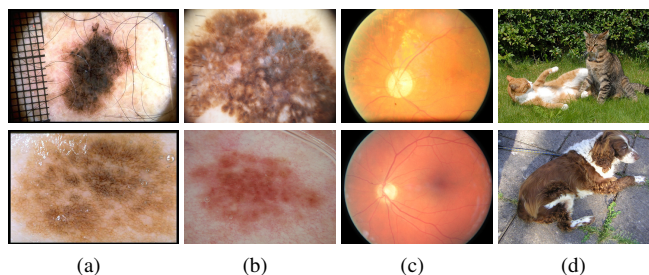
**Index Terms**— Melanoma screening, dermoscopy, deep learning, transfer learning

## 1. INTRODUCTION

From all skin cancers, melanoma represents just 1% of cases, but 75% of deaths<sup>1</sup>. Melanoma’s prognosis is good when detected early, but deteriorates fast as the disease progresses. Automated tools may play an important role in providing timely screening, helping doctors focus on patients or lesions at risk, but the disease’s characteristics — rarity, lethality, fast progression, and diagnosis subtlety — make automated screening for melanoma particularly challenging.

Vast literature exists on automated melanoma screening [1], but only in the past two years Deep Neural Networks (DNNs) [2] were proposed to improve accuracies [1, 3, 4, 5, 6, 7, 8, 9, 10, 11]. DNNs are the state of the art for image classification, but their use for medical images is challenging, since those models require very large training sets (from dozens of thousands, up to several million images) [12]. To bypass that difficulty, most current literature employs *transfer learning*, a technique where a model trained for a given *source task* is partially “recycled” for a new *target task*. Transfer learning ranges from simply using the output of the source DNNs as a feature vector, and training a completely new model (e.g., an SVM) for the target task; until using a pre-trained source DNN to initialize some of the weights of the target DNN, and training the latter as usual [13].

Works employing DNNs for melanoma screening either train a network from scratch [6, 7, 9], or transfer knowledge from ImageNet [4, 5, 8, 10, 11]. The main difference between works is



**Fig. 1.** Samples from datasets used here: (a) Atlas; (b) ISIC; (c) Retinopathy; (d) ImageNet. Each row shows a sample from a different class in the dataset. In this paper, datasets *c* and *d* are source datasets used for transferring knowledge to target models trained in the target task of melanoma screening, trained and evaluated in datasets *a* and *b*.

the choice of DNN architecture and implementation framework — the most common framework is Caffe [4, 8, 11], and the most common architectures are ResNet [10], DRN-101 [11], AlexNet [5], and VGG-16 [8]. Schemes for artificially augmenting training data, and for transferring learned knowledge also vary.

Segmenting the lesion before feature extraction is a common choice for melanoma screening [8, 11], although DNNs generally allow forgoing it [1, 4, 5, 8, 10, 14]. In this paper we forgo all pre-processing steps, like lesion segmentation, or artifacts/hair removal.

Because transfer learning seems so central for successful application of DNNs in melanoma screening, the main contribution of this paper is studying it in more detail. Figure 1 illustrates both the source datasets (Retinopathy and ImageNet) and the target datasets used in our experiments. We attempt to answer the following questions: (1) What is the impact of transfer learning? Current art is already convinced it helps, but by how much? (2) Is it better transferring from a related (medical) but smaller dataset, from a larger unrelated (general) dataset, or from both? (3) Does retraining the transferred network for the new task (known in the DNN literature as *fine tuning*) help, and by how much? Our aim in this paper, thus, is less pushing the envelope on model accuracies, and more increasing our understanding on transfer learning. However, in the spirit of showing that the alternatives evaluated are realistic, we contrast our models to current art and show that they have similar performance.

Dermoscopic examination is complex, involving many types of lesions, although automated screening is usually advanced as a binary decision, with melanoma as the positive class, and all other lesions as the negative class. However, other skin cancers — notably basal cell carcinomas — become a challenge for that model: should

<sup>‡</sup>Those authors contributed equally.

<sup>\*</sup>Corresponding author: dovalle@dca.fee.unicamp.br.

<sup>1</sup>American Cancer Society: <http://www.cancer.org>

they just be put in the negative class with all other lesions? should they be considered together with melanoma in a wholesale “malignant” class? should they be removed altogether from the analysis? or should they be analyzed as a third separate class? A second contribution of this paper is evaluating the impact of such decisions.

## 2. DATA AND METHODS

A reference implementation with all source code for this paper is available<sup>2</sup>. All software libraries mentioned here are linked there.

The datasets employed to train and test the target models (melanoma screening) were the Interactive Atlas of Dermoscopy (Atlas) [15], and the ISBI Challenge 2016 / ISIC Skin Lesion Analysis towards Melanoma Detection — Part 3: Disease Classification (ISIC) [16]. Both are available, the former commercially, and the latter free of charge.

The Interactive Atlas is a multimedia guide (Booklet + CD-ROM) designed for training medical personnel to diagnose skin lesions. The CD-ROM contains 1,000+ clinical cases, each with at least two images of the lesion: close-up clinical image, and dermoscopic image. Most images are 768 pixels wide  $\times$  512 high. Besides the images, each case is composed of clinical data, histopathological results, diagnosis, and level of difficulty. The latter measures how difficult (low, medium and high) the case is considered to diagnose by a trained human. The diagnoses include, besides melanoma (several subtypes), basal cell carcinoma, blue nevus, Clark’s nevus, combined nevus, congenital nevus, dermal nevus, dermatofibroma, lentigo, melanosis, recurrent nevus, Reed nevus, seborrheic keratosis, and vascular lesions. There is also a small number of cases classified simply as ‘miscellaneous’. In this paper we employed only the dermoscopic images, and included all cases except the acral lesions (whose particular appearance poses very specific challenges). Some images contained a black rectangular “frame” around the picture, which we cropped automatically with ImageMagick<sup>3</sup>.

The ISBI Challenge 2016 / ISIC Skin Lesion Analysis towards Melanoma Detection — Part 3: Disease Classification [16] is a subset of 1,279 dermoscopy images from the *International Skin Imaging Collaboration*, an international effort to improve melanoma diagnosis. The Challenge Dataset contains 900 images for training (273 being melanomas) and 379 for testing (115 being melanomas). Ground truth lesion segmentations are available, but we did not use them.

The sources datasets employed for the transfer (pre-training of the DNNs) were the Kaggle Challenge for Diabetic Retinopathy Detection dataset (Retinopathy)<sup>4</sup>, with a training set of 35,000+ high-resolution retina images taken under varying conditions; and the ImageNet Large Scale Visual Recognition Challenge 2012 dataset (ImageNet), containing 1M training images labeled into 1,000 categories [3]. Both datasets are freely available. Our use of ImageNet dataset was indirect — because training over it is so time consuming, we opted for employing pre-trained source DNNs.

Figure 1 illustrates those four datasets.

With a single exception, all protocols were based upon the VGG-M model proposed by Chatfield et al. [17]. We also run a single comparison with the VGG-16 model [18] to evaluate the impact of that deeper (and more expensive) architecture. Because the original implementations are in MATLAB, which was not convenient for our computational infrastructure, we reimplemented those models in

Lasagne<sup>5</sup> and Nolearn<sup>6</sup>.

In the experiments with transfer learning, we get the source networks pre-trained on ImageNet, train them from scratch on Retinopathy, or fine-tune on Retinopathy the model pre-trained on ImageNet. In the baseline (control) experiment without transfer learning, the networks are trained from scratch. In all networks, we ignore the output layer and employ an SVM classifier to make the decision. We did so for all experiments, including the fine-tuned ones, to avoid introducing extraneous variability.

Whenever training is involved (when we fine-tune or train networks from scratch) we employ the technique of *data augmentation*, which consists in creating randomly modified training samples from the existing ones. Data augmentation is almost always helpful [2, 17, 19], but particularly important when working with the small datasets typical of medical tasks. Augmentation also brings the opportunity to rebalance the classes, alleviating another problem of medical tasks, where the positive class (melanomas) is usually much smaller than the negative (benign lesions). To accomplish both augmentation and balance, we only augment the minority classes (Melanoma, Malignant, and Basal Cell Carcinoma, depending on the experimental design).

The factors investigated are: *transfer learning*, which can be *no transfer* (Atlas from scratch), *from Retinopathy* (Atlas from Retinopathy), *from ImageNet* (Atlas from ImageNet), or *double* (Atlas from Retinopathy from ImageNet); and *fine tuning*, which can be *no FT* (only the final SVM decision layer is trained in the target models), or *with FT* (the target model is retrained for melanoma, before the last layer is stripped and the SVM decision layer is trained).

The protocol, in detail:

1. The weights of the networks are initialized as follows: (a) For experiments without transfer, the weights of the network are initialized as an orthogonal matrix while the biases are initialized as constants (0.05); (b) For experiments transferring from ImageNet, the weights are loaded from an available pre-trained network<sup>7</sup> (converting from MATLAB format to Nolearn format); (c) For experiments transferring from Retinopathy, the weights are initialized as 1.a, and the network is trained for Retinopathy; (d) For experiments with double transfer, the weights are loaded from ImageNet as explained in 1.b, and the network is fine-tuned for Retinopathy;
2. To reach the input size required by VGG, we resize all images to  $224 \times 224$  pixels, distorting the aspect ratio to fit when needed;
3. We “re-center” the input, as required by VGG, by subtracting the mean of the training dataset. For networks transferred from ImageNet, we used the (precalculated) mean that comes with VGG; for the others, we calculated the mean of the images from the respective training sets;
4. (Only for the protocols involving DNN training, either from scratch, or with fine-tuning:) We strip the last layer of the source network, add a new softmax layer to make the decision on the skin lesion, and retrain the entire network. When training or fine-tuning the model for Retinopathy, we proceed just like training for Melanoma, except that output softmax units take 5 classes (as specified in Kaggle Competition) instead of 2 or 3 for Melanoma, or 1,000 for ImageNet. We balance the training set by augmenting the data for the minority classes, applying a

<sup>2</sup><https://sites.google.com/site/robustmelanomascreeing>

<sup>3</sup><http://www.imagemagick.org>

<sup>4</sup><https://www.kaggle.com/c/diabetic-retinopathy-detection/data>

<sup>5</sup><https://lasagne.readthedocs.io/en/latest/>

<sup>6</sup><http://nolearndocs.readthedocs.io/en/latest/lasagne.html>

<sup>7</sup><http://www.vlfeat.org/matconvnet/pretrained/>

**Table 1.** Main results (AUC in %; FT: fine tuning). Surprisingly, transfer from another specific medical task (Retinopathy) is not effective, even if preceded by transfer from a general task (ImageNet). Fine-tuning has major impact and should be considered a necessity. The choice of labeling has a small and somewhat inconsistent impact, that might be due to chance.

Experimental Design	No Transfer	From Retinopathy		From ImageNet		Double Transfer with FT
		no FT	with FT	no FT	with FT	
Malignant vs. Benign	76.0	72.8	76.0	79.1	<b>82.5</b>	78.8
Melanoma vs. Benign	75.7	73.5	75.3	77.9	<b>80.9</b>	<b>80.9</b>
Melanoma vs. Carcinoma vs. Benign	73.0	71.4	72.8	79.4	<b>83.6</b>	81.8

**Table 2.** Results on ISIC, with VGG-16, with transfer learning from ImageNet, with fine tuning. Baselines quoted from the ISIC 2016 competition website (competitors were originally ranked by mAP). The results are provided to show that the models evaluated here are realistic — in the sense that they are in the same ballpark of performance as current art. All numbers in %.

	AUC	mAP	ACC	SE	SP
1 <sup>st</sup> place	80.4	63.7	85.5	50.7	94.1
2 <sup>nd</sup> place	80.2	61.9	83.1	57.3	87.2
3 <sup>rd</sup> place	82.6	59.8	83.4	32.0	96.1
This paper	80.7	54.9	79.2	47.6	88.1

random transformation (scaling, rotating, or flipping the images). The random number generator is fixed at the very beginning of each experiment, so that the training sequence is the same for all experiments. That makes the comparisons fairer, reducing uncontrolled variability. We train for 60 epochs with learning rate  $10^{-3}$ , Nesterov momentum 0.1, and  $\ell_2$ -regularization of  $5 \times 10^{-3}$ . We perform a 10% stratified split of the training set for validation. We keep the model that minimizes validation loss;

- We extract the outputs of Group 7/Layer 19, which are vectors of 4,096 dimensions, and use that output as a feature-vector to describe the images; we  $\ell_2$ -normalize those vectors, and feed them to a linear SVM classifier (Sklearn implementation) to make the decision on the skin lesion. We separate 10% of the training set for validation, to choose the SVM regularization parameter, optimizing on the AUC.

We evaluated three experimental designs, varying the labeling of the classes:

- Malignant vs. Benign lesions: melanomas and basal cell carcinomas were considered positive cases and all other diagnoses were negative cases;
- Melanoma vs. Benign lesions: melanomas were positive cases while all other diagnoses were negative ones, removing basal cell carcinomas;
- Basal cell carcinoma vs. Melanoma vs. Benign lesions: here we have three classes, with all other diagnoses under a single Benign label.

For all designs we employed 5×2-fold cross-validation. Our splits were semi-random, making an effort to balance as much as possible diagnose distributions, to avoid unnecessary variability.

Our main metric was the Area Under the ROC Curve (AUC); for the design with three classes, we computed three one-vs-one AUCs and reported their average. For the experiment with the ISIC dataset we also report the other measures employed in the competition. We show the results on ISIC for reference purposes, to demonstrate that the models being discussed here are in the same ballpark of performance as the current state of the art (Table 2).

**Table 3.** Impact of the DNN architecture choice. A deeper model (VGG-16) leads to best results, regardless of the experimental design. All experiments with transfer from ImageNet and fine tuning.

Architecture	AUC (%)		
	Mal×Ben	Mela×Ben	Mela×Carc×Ben
VGG-M	82.5	80.9	83.6
VGG-16	<b>83.8</b>	<b>83.5</b>	<b>84.5</b>

**Table 4.** Results stratified by diagnosis difficulty of test images (Low, Medium or High), for VGG-M, transferring from ImageNet, with fine tuning. All: performance over the whole dataset. Low-, medium-, and high- difficulty cases represent respectively 38.1, 36.3, and 25.6% of the whole dataset.

Experimental Design	AUC (%)			
	Low	Medium	High	All
Malignant vs. Benign	93.7	82.5	58.8	82.5
Melanoma vs. Benign	93.0	79.6	56.6	80.9

### 3. RESULTS

Our results are in Table 1. Fine-tuning improves classification, both when transferring from the small-but-related dataset (Retinopathy), and when transferring from the large-but-unrelated task (ImageNet): that agrees with current literature in DNNs, which almost always endorses fine-tuning. Surprisingly, transfer learning from Retinopathy (also a medical-image classification task) leads to worse results than transferring from the general task of ImageNet, even in combination with the latter. That might indicate that transferring from very specific tasks poses special challenges for overcoming the specialization — even if the source and target tasks are somewhat related. The best protocol we found was to simply transfer from ImageNet, with fine-tuning. The comparison between DNN architectures shows that — as usually observed for image classification — a deeper DNN performs better (Table 3).

The experimental designs also showed differences in performance: in general it was easier to either group Basal cell carcinomas with Melanomas (Malignant vs. Benign), or to consider them as a separate class (Melanoma vs. Carcinomas vs. Benign), than to ignore them altogether (Melanoma vs. Benign). Those results suggest that organizing the labels affects the difficulty of the task, but the explanation for those aggregate numbers might be simply that Basal cell carcinomas are easier to diagnose than Melanomas.

We show the results stratified by diagnose difficulty (as indicated by the Atlas itself) in Table 4. Those results show that low-difficult lesions can essentially be solved by current art with relatively high confidence, while for difficult lesions performance is still little better than chance.

## 4. CONCLUSIONS

Our results are consistent with current art on DNNs: transfer learning is a good idea, as is fine tuning. Our results also suggest, in line with literature in DNNs, that deeper models lead to better results. We expected that transfer learning from a related task (in our case, from Retinopathy, another medical classification task) would lead to better results, especially in the *double* transfer scheme, that had access to all information from ImageNet as well. The results showed the opposite, suggesting that adaptation from very specific — even if related — tasks poses specific challenges. Still, we believe that further investigation is needed (e.g., can another medical task show better results? can another transfer scheme work?).

The results suggest that the experimental design is sensitive to the choice of lesions to compose the positive and negative classes, maybe due to the relative difficulty of identifying each of the types of cancer evaluated (Melanomas and Carcinomas).

The results stratified by diagnose difficulty suggest that current art can already deal with the lower and middle spectrum of difficulty, especially considering that human doctors' accuracies might be between 75-84% [20]. On the other hand, difficult lesions appear *really* hard to diagnose. That suggests a *referability* framework as potentially more fruitful than a *diagnostics* framework. Referring to the doctor both the cases in which the model has high confidence for the positive label, and the *hard* cases (for which the model has low confidence), might be more achievable in the short term than attempting to have high confidence for all cases.

## 5. ACKNOWLEDGEMENTS

We are very grateful to the supporters that made this research possible. E. Valle is supported by CNPq/PQ grant 311486/2014-2. M. Fornaciali, R. Pires, and E. Valle are partially funded by Google Research Awards for Latin America 2016. A. Menegola is funded by CNPq; R. Pires is partially funded by CAPES; S. Avila is funded by PNPd/CAPES. We also acknowledge the support of CNPq (Universal grants 2016); and NVIDIA Corporation, with the donation of a Tesla K40 GPU used for this research. We also thank Prof. Dr. M. Emre Celebi for kindly providing the machine-readable metadata for The Interactive Atlas of Dermoscopy.

## 6. REFERENCES

- [1] M. Fornaciali, M. Carvalho, F. Vasques Bittencourt, S. Avila, and E. Valle, "Towards automated melanoma screening: Proper computer vision & reliable results," *arXiv preprint arXiv:1604.04024*, 2016.
- [2] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
- [4] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. Smith, "Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images," in *MLMI*, 2015, pp. 118–126.
- [5] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *ISBI*, 2016, pp. 1397–1400.
- [6] A. Masood, A. Al-Jumaily, and K. Anam, "Self-supervised learning model for skin cancer diagnosis," in *EMBS NER*, 2015, pp. 1012–1015.
- [7] J. Premaladha and K. Ravichandran, "Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms," *J Med Syst*, vol. 40, pp. 1–12, 2016.
- [8] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *ECCV*, 2016, pp. 206–222.
- [9] S. Demyanov, R. Chakravorty, M. Abedini, A. Halpern, and R. Garnavi, "Classification of dermoscopy patterns using deep convolutional neural networks," in *ISBI*, 2016, pp. 364–368.
- [10] F. Cicero, A. Oliveira, and G. Botelho, "Deep learning and convolutional neural networks in the aid of the classification of melanoma," in *SIBGRAPI*, 2016.
- [11] N. Codella, Q.-B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, and J. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J Res Dev*, vol. 61, 2017.
- [12] H.-C. Shin, H. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *TMI*, vol. 35, no. 5, pp. 1285–1298, 2016.
- [13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *NIPS*, 2014, pp. 3320–3328.
- [14] M. Fornaciali, S. Avila, M. Carvalho, and E. Valle, "Statistical learning approach for robust melanoma screening," in *SIBGRAPI*, 2014, pp. 319–326.
- [15] G. Argenziano, H. Soyer, V. Giorgi, D. Piccolo, P. Carli, M. Delfino, et al., "Dermoscopy: a tutorial," *EDRA, Medical Publishing & New Media*, 2002.
- [16] D. Gutman, N. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint arXiv:1605.01397*, 2016.
- [17] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.
- [18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [19] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *CVPR Workshops*, 2014, pp. 806–813.
- [20] A.-R. Ali and T. Deserno, "A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data," in *Proc. SPIE 8318, Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*, 2012, vol. 8318, pp. 831811–1.