INTRAPATIENT DIAGNOSIS FOR MELANOMA SKIN CANCER

António Mendes *Catarina Barata*

Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal

ABSTRACT

Melanoma is the most fatal form of skin cancer, making its research a hot topic worldwide. Despite the promising results achieved by recent methods, they often disregard one of the most important criteria considered by dermatologists: the intrapatient lesion context. In this work we address this limitation, leveraging the medical concept of the ugly duckling (UD) lesion, and translating it to both the deep learning and the anomaly detection fields. Considering this, we propose an integrated diagnosis at the image and patient levels. This model aggregates blocks of supervised trained networks, such as Convolutional Neural Networks and self-supervised learning as is the case of Autoencoders. Overall, our results strongly support the hypothesis that the intrapatient context can enhance the diagnosis. The detection of melanoma at the image level improved by 7%, while the patient diagnosis improved by over 10% compared with models that did not incorporate the patient context.

***Index Terms*—** Deep Learning, Outlier Detection, Melanoma, Ugly Duckling, Dermoscopy

1. INTRODUCTION

The rapid development of metastatic disease is one key reason for melanoma being the most fatal form of skin cancer. It is crucial to detect melanomas in the early stages to increase the likelihood of a successful treatment [1]. One way to support clinicians is through the development of automatic diagnostic systems.

Previous works have already proven the efficacy of deep neural networks (DNNs) for skin lesion classification using dermoscopy images [2], [3]. However, their output is often based, exclusively, on a single image input. Such approaches perform a greedy analysis of the patient, where each of its lesions is independently analyzed and it is assumed that no information can be gained from comparing them. During our experiments we observed that such assumptions can be misleading, resulting in poor diagnostic performances at the patient level. As an example, one of our models that achieved an area under the curve (AUC) above 87.0% at the image level, would **miss 40% of the melanoma patients and misdiagnose approximately 45% of the benign patients**, resulting in a high number of unnecessary excisions. These results raise the research community's awareness for the importance of incorporating a patient-level decision on the diagnostic systems.

In fact, dermatologists follow several standard methodologies during a screening session, opposed to a single image diagnosis. One evidence accepted inside the community to differentiate melanomas from benign nevus is the **ugly duckling (UD)** – usually described as an odd lesion that looks different from the other patient's lesions [4]. From a machine learning perspective, an UD can be seen as an outlier or anomaly that deviates from the normal phenotypic expression of a patient. The concept of UD has been recently explored by M. Mohseni et al. [5] and L. Soenksen et al. [6], who proposed multi-stage pipelines to detect outlier lesions using total body photography (TBP). TBPs are significantly different from dermoscopy as the first ones include large body parts and several lesions in the same image, which also implies less resolution. Both works approach the UD detection problem by computing anomaly scores for each lesion obtained from a segmentation process.

To the best of our knowledge, the only model addressing the detection of UD in dermoscopy images was proposed by Z. Yu et al. [7]. This is an end-to-end approach that passes the images through a convolutional neural network (CNN) and collects the deep features. Then, a transformer encoder receives these features and models the dependency between different lesions from the same patient. Finally, these embeddings feed a classification network. While this approach showcases the potential of UD to improve melanoma diagnosis, the use of transformers constrains the number of images that can be used to capture the patient context, meaning that for patients with a high number of lesions a few must be discarded.

In our work we aim to improve the melanoma skin cancer diagnosis by proposing a new strategy to incorporate the intrapatient context in DNNs. We propose an **integrated diagnosis** that comprises two main branches – an **image diagnosis and an auxiliary patient diagnosis**. In the end we merge them to output a diagnosis for each image that is influenced by the other lesions from the same patient. Briefly, the image diagnosis branch uses pretrained CNNs, and the

patient diagnosis consists of a logistic regression (LR) that receives patient embeddings. Those are built with preprocessed features from the CNN and reconstruction errors from a Convolutional Autoencoder (CAE) and take inspiration from traditional outlier detection strategies. Contrary to [7], our formulation does not assume a fixed number of lesions per patient. We experimentally demonstrate the validity of our approach, including a final generalization analysis using the held-out test set of ISIC 2020. This is, to the best of our knowledge, the first demonstration that intrapatient context can improve the performance of melanoma diagnosis across datasets.

The remain of the paper is organized as follows. The next section has the details regarding our experiments, focusing on the integrated diagnosis and the main reasonings for each block. Then, on section 3 we describe the experimental setup, which includes online data augmentation, dataset partition, and computational hardware. Section 4, reports and interprets the results obtained. Finally, in section 5, we highlight the major contribution of this work.

2. METHODOLOGIES

As declared before we are interested in optimizing the melanoma diagnosis given the characteristics of each patient. The idea is to search for the UD evidence, which can be interpreted as outlier samples when confronted with the patient's patterns of normality, i.e., the patient phenotype.

2.1. Dataset

The dataset comes from the 2020 SIIM-ISIC Melanoma Classification Challenge [4]. It comprises 32,701 images from over 2,000 patients and correspondent metadata, including encrypted patient identification, age group, sex, lesion's anatomic site, and lesion diagnosis. One main particularity of this dataset is the class imbalance. We considered two classes – the benign class with 98.2% of the total examples and the melanoma class with the 1.8% left. Also, the number of lesions per patient varies between two and 115. Furthermore, as the images come from various clinics, we created a pre-processing protocol to minimize disparities and standardize the input across the experiments. We apply the Shades of Gray algorithm [8] to correct the image's color and resized the output to 300x300 pixels using padding to maintain the aspect ratio. We excluded the 425 duplicated images listed in the challenge webpage [9].

2.2. Integrated Diagnosis

The integrated diagnosis is composed by an upper path responsible for the single image diagnosis, and a lower path for the patient diagnosis. Then, both paths are concatenated to feed a fully connected layer that is responsible for weighing the contribution of each path in the final diagnosis.

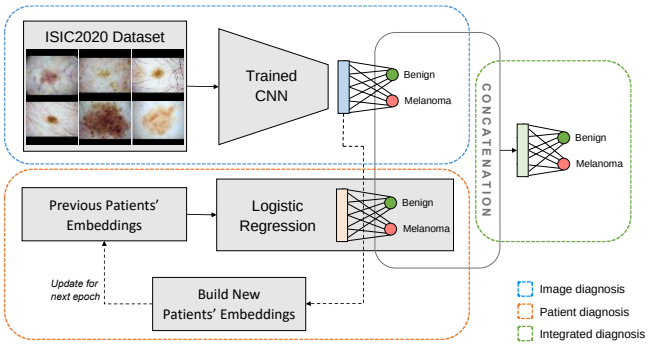

Fig. 2. Pipeline for the integrated diagnosis combining both the image and the patient diagnosis.

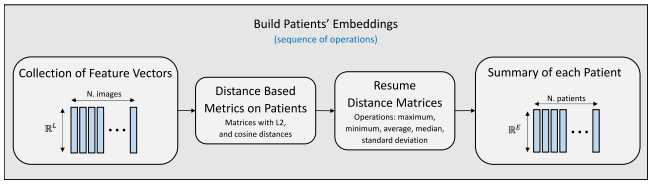

Fig. 3. From feature vectors to patient embeddings.

This pipeline is shown in Fig. 2. It is noteworthy that the final output scores already carry the intrapatient context.

Upper Path: This path is responsible for processing each dermoscopy image independently and producing a preliminary image level diagnosis. It is composed by a standard CNN architecture with a global average pooling (GAP) layer between the last convolutional layer and the output one. The output of the GAP for each image of a patient $z_p^n \in \mathbb{R}^L$ is also used by the lower path as described below.

Lower Path: This path performs a patient level diagnosis based on two types of descriptors. The first one relies on the comparison between the feature vectors z_p^{i} of the various images. Since we want to detect outliers (the UDs) in the images, we opted to compute Euclidean and cosine distances between the vectors of features and resume these high dimension distance matrices using the five statistical operations summarized in Fig. 3. The second descriptor leverages the recent advances with self-supervised learning (SSL) and trains a Convolutional Autoencoder (CAE). This CAE was optimized to reconstruct the dermoscopy images, minimizing the mean squared error (MSE). We believe this may be a great way of taking advantage from the dataset imbalance, as the CAE has less chances to learn from the melanoma examples. Consequently, we expect discrepancies in the reconstruction performance for each class. The CAE is applied to all images of a patient and the MSEs are computed. Finally, the errors are aggregated as in the case of the GAP vectors, using the five statistical operations in Fig. 3. The descriptor's combination is called a patient embedding, $h_p \in \mathbb{R}^E$. The patient embedding is then fed to a LR that predicts if a patient has at least one melanoma.

3. EXPERIMENTAL SETUP

3.1. Model Training

At training time, we used a warm start initialization whenever possible. This means that we start by training each path independently and only then train the integrated model. For the upper path, we pretrain a CNN on the ISIC 2020 set, using a weighted binary cross-entropy (BCE) loss to predict a diagnosis for each image. We penalize errors in the benign class with a weight $w_b = 0.02$, and $w_m = 0.98$ for the melanoma class. The choice aims to reflect an approximation to the classes' prior probability. This strategy prevents the converge to solutions in which the classifier is biased towards the major class. The LR was pretrained using patient embeddings collected from the previous described CNN as input. In this case we predict if a patient has at least one melanoma among its lesions. For that reason, we adjusted the BCE weights to $w_b = 0.2$ and $w_m = 0.8$. The integrated diagnosis is trained as follows. On each epoch, we train the upper path and the weights after the concatenation operation, while keeping the lower path with its weights frozen. Additionally, before the first epoch and on every $K \in \mathbb{N}$ epochs, we compute the new embeddings for each patient, with the features collected from the current upper path model, replicating the process shown in Fig. 3. After the embeddings update, we train the LR, exclusively, until notice convergence or reaching a maximum number of epochs. Then, in case the new patient diagnosis performs better than the previous, we update the general model with this best new model. This way we distribute the training process to alternately optimize both paths. Regarding the loss functions used, we kept the loss function from the warm training. The only modification made is related with the logits normalization before computing the losses. Similarly to the authors of [10] we normalize the logits to avoid overconfident diagnosis' probabilities. The normalized logits, ℓ', are obtained by the following operation,

$$\ell' = \frac{\ell}{T \cdot (\|\ell\|_2 + \zeta)} \ , \tag{1}$$

where ℓ represents the logits, in this case, vectors of two components, and T the temperature constant that scales the normalized logits magnitude. In this work we use $T = 0.005$ as suggested in [10]. The constant $\zeta = 10^{-7}$ was used to ensure numerical stability.

One way to improve the model generalization to new data is by having a more diverse dataset. In our case, we applied a set of online transformations to the images at training time. Specifically, random horizontal and vertical flips, both with an independent probability of 50%, followed by a random erasing. This last transformation replaces a rectangle of the image with black color pixels. The frequency of occurrence, and the rectangles' size and ratio were tuned as suggested in [11]. The only experiment that didn't use this transformation is the CAE for lesion reconstruction since we did not want it to learn so much noise. Additionally, when using the ImageNet initialization, we normalized each image channel with the mean and standard deviation of the ImageNet dataset as claimed in [12]. All experiments were performed on a computer equipped with an Intel(R) Core(TM) i7-7700 CPU, 16GB RAM, and an NVIDIA GeForce GTX 1060 GPU, 6GB.

3.2. Evaluation Strategy and Metrics

We split the ISIC 2020 train set to have a set of unseen data. The partition was random, while simultaneously respecting the following criteria: i) the training and validation sets are disjointed with proportions of 80% and 20%, respectively; ii) all lesions from a patient need to be grouped, either in the training or the validation sets. In the end, both sets maintained the original proportions between classes. Furthermore, the ISIC 2020 challenge provides a test set, but the ground truth labels are not publicly available. So, the test set can only be used to assess the model's performance through the challenge submission site [13]. We performed a final and unbiased evaluation of our models with this set. To evaluate the results, we computed the following metrics: i) specificity (SP); ii) sensitivity (SE); iii) balanced accuracy (BAC); and AUC.

4. RESULTS

We tested several combinations of popular CNN architectures for the upper path and the encoder of the CAE. We report the results for the best configurations, however additional models can be accessed in our GitHub[1]. For the trained CNN we picked the EfficientNet-b2. For the LR we opted for the embeddings resulting from the EfficientNet-b2, and the U-Net Resnet18 (CAE) reconstruction errors.

Table 1 shows the results for our validation and ISIC test sets, both at the image and patient level. The baseline model corresponds to training only the upper path in Fig. 2. We also report the performance for two integrated models: in the first one only EfficientNet-b2 is used to build the patient embeddings to the lower path, while in the second we also consider the CAE reconstruction errors. The performance per patient is assessed by converting the results per image into a patient diagnosis as follows: if at least one of the images is diagnosed as melanoma, then the patient will be diagnosed accordingly. While we could have used the lower path for this, performing a final patient diagnosis based on the images is a reasonable assumption that matches the clinical practice and allows us to compare our results with the baseline.

Overall, our results suggest a performance increase when the patient context is used in the diagnosis. The integrated diagnosis has a greater melanoma SE in both performances' assessments, per image (~7%) and per patient (~10%). In fact, the biggest advantage of the integrated diagnosis is the ability to improve the intrapatient overall diagnosis.

[1] Upon acceptance of the paper, the code will be released at github.com/antoniogamamendes.

Table 1. Results evaluated per image and per patient considering the baseline diagnosis and the integrated diagnosis (without and with the CAE reconstruction errors). The best BAC's are in bold.

	Network	Metric	Validation	Test
Performance Per Image	Baseline (EfficientNet-b2)	SP	0.918	0.916
		SE	0.607	0.552
		BAC	0.763	0.734
		AUC	0.870	0.858
	Integrated Diagnosis (EfficientNet-b2)	SP	0.920	0.916
		SE	0.598	0.585
		BAC	0.759	**0.751**
		AUC	0.808	0.767
	Integrated Diagnosis (EfficientNet-b2 + U-Net Resnet-18)	SP	0.902	0.895
		SE	0.650	0.590
		BAC	**0.776**	0.743
		AUC	0.848	0.805
Performance Per Patient	Baseline (EfficientNet-b2)	SP	0.554	-
		SE	0.602	
		BAC	0.578	
	Integrated Diagnosis (EfficientNet-b2)	SP	0.548	
		SE	0.830	
		BAC	**0.689**	
	Integrated Diagnosis (EfficientNet-b2 + U-Net Resnet-18)	SP	0.545	
		SE	0.830	
		BAC	0.687	

Fig. 5. Lesions from a 90 year old male patient. On the left, the baseline diagnosis gave the right diagnosis for one of the melanomas, producing a FP and a FN. As a result of the integrated diagnosis, the FN diagnosis has been corrected, however, the label for the first image remains incorrect. The ground truth label (text) and the predicted one (image contour – red melanoma and green benign) are reported.

Furthermore, the integrated diagnosis benefits from combining the CAE reconstruction errors statistical operations in the patient embeddings. As consequence, when comparing both integrated diagnosis, with and without the CAE, the AUC increased by 5%. Finally, the test set performance did not deviate too much from the validation set, which indicates developed the models did generalize well, at least for this particular dataset. Our scores are also in line with what is reported in the submission platform, where the AUC scores' range for the test set is [0.472, 0.946]. It is important to stress that we do not perform additional steps to improve performance, such as test time augmentation or ensemble. To translate this numerical overview, we present from Fig. 5 to Fig. 7, three clinical examples of patients for whom integrating their context produced a better diagnosis outcome than the one achieved by the baseline diagnosis (without context). For simplicity, we choose some scenarios where the integrated diagnosis helps. Meaning that corrections were made in different directions (FN and FP samples).

A direct comparison between our results and those reported in [7] is not possible, since different partitions of the

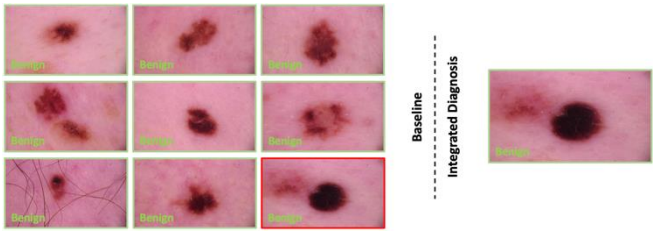

Fig. 6. Nine lesions from a 55 year old male patient. In this case, all lesions were benign, but the baseline diagnosis produced a FP. The mistake was corrected in the integrated diagnosis that ended up giving a full correct diagnosis for all the patient's lesions. The ground truth label (text) and the predicted one (image contour – red melanoma and green benign) are reported.

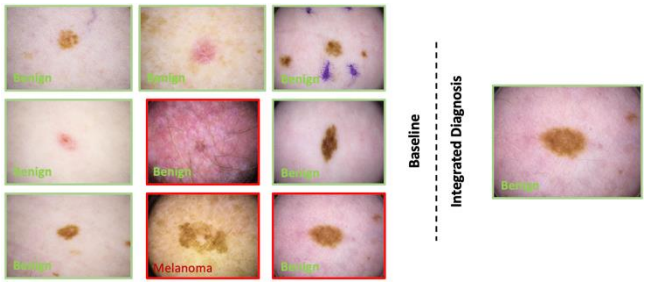

Fig. 7. Nine lesions from a 35 year old woman. Despite being able to detect only melanoma, the baseline diagnosis also produced two FP diagnoses. However, the patient context helped the integrated diagnosis discard one of the mistakes. The FP not corrected may be a difficult assessment due to the significant colorization difference to the other lesions. The ground truth label (text) and the predicted one (image contour – red melanoma and green benign) are reported.

ISIC 2020 are used for training and validation, and they do not assess the performance on the test set. Nevertheless, we are able to achieve similar improvements in melanoma SE (~6% for the full model in [7] vs ~7% in this work) without constraining the number of images per patient. Additionally, we achieve greater improvements in the diagnosis at the patient level (~10% in our case, vs ~2% in [7]).

5. CONCLUSIONS

This work addressed the problem of melanoma diagnosis in scenarios where multiple dermoscopy images for the same patient are available. To address this challenge, we proposed an integrated diagnosis system that combines a diagnosis at the image and patient levels. The integrated system did successfully show a positive effect in combining several images from the same patient to improve the overall diagnosis. The detection of melanoma at the image level improved 7%, while the patient diagnosis improved by more than 10%, compared to the baseline models which did not include the patient context. These results open the door to further research on the use of patient context to improve the performance, safety, and trustworthiness of automatic diagnosis systems.

6. ACKNOLEDGMENTS

This research was funded by the project IntelligentCare – Intelligent Multimorbidity Managment System (Reference LISBOA-01-0247-FEDER-045948) is co-financed by the ERDF – European Regional Develpment Fund through the Lisbon Portugal Regional Operational Program – LISBOA 2020 and by the Portuguese Foundation for Science and Technology – FCT under CMU Portugal. It was also supported by the multi-year funding [LARSyS - FCT Plurianual funding 2020-2023. C. Barata is paid by the [CEECIND/ 00326/2017]. We thank Nicholas Kurtansky and Dr. Veronica Rotemberg for their insightful comments and discussion.

7. REFERENCES

[1] C. Gaudy-Marqueste *et al.*, "Ugly duckling sign as a major factor of efficiency in melanoma detection," *JAMA Dermatol*, vol. 153, no. 4, pp. 279–284, Apr. 2017.

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

[3] C. Barata and J. S. Marques, "Deep Learning For Skin Cancer Diagnosis With Hierarchical Architectures; Deep Learning For Skin Cancer Diagnosis With Hierarchical Architectures," *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, [Online].

[4] V. Rotemberg *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci Data 8, 34*. 2021.

[5] M. Mohseni, J. Yap, W. Yolland, A. Koochek, and S. Atkins, "Can self-training identify suspicious ugly duckling lesions?".

[6] L. R. Soenksen, T. Kassis, S. T. Conover, B. Marti-Fuster, J. S. Birkenfeld, and J. Tucker-Schwartz, "Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images," *José Avilés-Izquierdo*, vol. 13, p. 17, 2021, [Online].

[7] Z. Yu *et al.*, "End-to-End Ugly Duckling Sign Detection for Melanoma Identification with Transformers," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021, vol. 12907 LNCS, pp. 176–184. doi: 10.1007/978-3-030-87234-2_17.

[8] "Kaggle - Shades of Gray Algorithm." https://www.kaggle.com/code/maryadewunmi/isic-melanoma-classification (accessed Mar. 01, 2022).

[9] "Welcome to the ISIC Challenge." https://challenge.isic-archive.com (accessed Dec. 12, 2021).

[10] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating Neural Network Overconfidence with Logit Normalization," May 2022, [Online]. Available: http://arxiv.org/abs/2205.09310

[11] "Transposed convolution and checkerboard artifacts." https://distill.pub/2016/deconv-checkerboard/ (accessed May 18, 2022).

[12] "Image-Net Mean and Standard Deviation" https://github.com/developer0hye/PyTorch-ImageNet (accessed Mar. 15, 2022).

[13] "Kaggle challenge submission page (online)." https://www.kaggle.com/competitions/siim-isic-melanoma-classification/data (accessed May 17, 2022).