UGLY DUCKLING – OUTLIER DETECTION IN DERMOSCOPY IMAGES USING DEEP NEURAL NETWORKS

António Manuel de Almeida Gama Mendes
Instituto Superior Técnico, Universidade de Lisboa
Email: antonio.gama@tecnico.ulisboa.pt

ABSTRACT

We addressed the problem of the melanoma skin cancer diagnosis. Melanoma is the most fatal form of skin cancer, which makes this a topic of interest worldwide. Our goal is to improve diagnosis by incorporating the intrapatient context and contribute to this underexplored topic in the literature. We used the medical ugly duckling concept, supported by clinicians, as the foundation for transferring it to both the deep learning and the anomaly detection fields.

Considering this, we implemented an integrated diagnosis system that combines a diagnosis at both 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 the autoencoders and the generative adversarial networks.

Overall, results strongly support the hypothesis that intrapatient context can enhance diagnosis. Not only did the image diagnosis improve by over 2% in the balanced accuracy, but also the patient diagnosis by more than 10%, when compared to baseline models which did not include the patient context. Additionally, we developed a one-class classifier generative adversarial network for single lesion diagnosis. This model achieved a balanced accuracy of 82.6% in the validation set. Finally, we also discovered that the lesion's anatomic site does not appear to interfere with the diagnosis, contrarily to the images' hospital source.

***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 treatment success likelihood [1]. One way to support clinicians might be through automatic diagnosis systems.

Previous works' results have already proven the efficacy of deep neural networks (DNNs) for skin lesion classification using dermoscopy images [2], [3]. However, their output is based, exclusively, on a single image input. This issue opened the opportunity to research methods optimized for multi-lesion diagnosis, which rely on the patient context.

In fact, dermatologists also 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].

Recently, M. Mohseni *et al.* [5] and L. Soenksen *et al.* [6] 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 [5], [6] approach the UD detection problem by computing anomaly scores for each lesion obtained from a segmentation process. M. Mohseni *et al.* fine-tune a pretrained AE with lesions from a single patient. Then extracts the features for each lesion and computes their ℓ_2 norm, used after to define a threshold. Finally, the classification is done by comparing this threshold value with the AE reconstruction loss. L. Soenksen *et al.* extracts features from the final layer of a DNN trained for suspicious lesion detection. In parallel, these features are transformed using geometric distances and used as input to determine the UD lesions.

Furthermore, Z. Yu *et al.* proposes an end-to-end pipeline to detect UD lesions using dermoscopy images [7]. The approach passes the images through a convolutional neural network (CNN) and collects the deep features. Then, they use a transformer encoder that receives these features and models the dependency between different lesions from the same patient. Finally, these embeddings feed a classification network.

In our work we aim to improve the melanoma skin cancer diagnosis by incorporating the intrapatient context and see how DNNs handle the UD samples from each patient. For this reason, we propose an integrated diagnosis. It comprises two main branches – an image diagnosis and a 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 pre-trained

CNNs, and the patient diagnosis consists of a logistic regression that receives patient embeddings. Those are built with preprocessed features from the CNN and reconstruction errors from a convolutional autoencoder (CAE).

As a research complement, we also tried to develop a one-class generative adversarial network (GAN). This idea came from a work [8] that is not related to the skin cancer diagnosis, but comes from the outlier detection field. So, we found it had potential and decided to experiment transfer the concept to the melanoma diagnosis. The framework described in [8] consists of training a GAN with examples from a class. In this stage the discriminator tries to distinguish between the real examples and the ones artificially created by the generator. Then, at training time, we show examples from different classes, expecting that the GAN recognizes them as fake examples. Analogously, we can explore the same strategy to distinguish melanomas from benign lesions.

The next section presents the methodology followed in this work, closing with the integrated diagnosis and the main reasonings for each block. Then, on sections 3 we describe the experimental setup, which includes the online data augmentation, the dataset partition, and the computational hardware. Section 4, reports and interprets the results obtained across the different experiments. Finally, in section 5, we highlight the major contribution of this work and potential future work directions.

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 confronted with the patient's patterns of normality.

2.1. Dataset and pre-processing

The data source is 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[1]. 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, the images come from five clinics – New York, Barcelona, Brisbane, Vienna, and Sydney. So, we created a pre-processing protocol shown in Fig. 1 to minimize disparities and standardize the input across the experiments. We apply the Shades of Gray algorithm [9] to correct the images' color and resized them to 300x300 pixels using

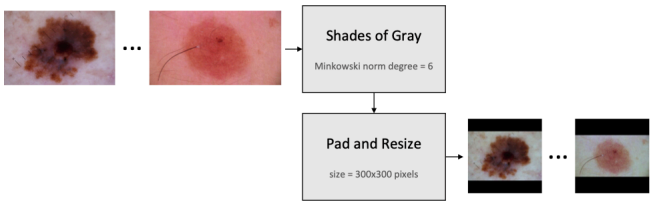

Fig. 1. Images before (left) and after (right) pre-processing.

padding to maintain the aspect ratio and avoid images' distortion or alternatively do center crops, which may unintentionally remove important parts from the lesions.

2.2. Single image diagnosis

The first stage of our work consists of building a single image diagnosis that do not uses the patient context. We call it our baseline diagnosis and there are two main reasons to do so. One is to have a benchmark performance without context to later understand if introducing the context is advantageous. Second, after training these networks, we can use them as warm starts and integrate them into our main pipeline. In this first classification task we tested six different architectures (EfficientNet-b1, EfficientNet-b2, Resnet-18, Resnet-34, Resnet-50, and Inception-v3). All were initialized with the ImageNet weights and the output layer was adjusted to have just two neurons instead of the original 1,000 as our classes of interest are the benign and melanoma sets. Additionally, we have added a dropout layer with probability equal to 0.3 to the Resnet networks, since they did not originally have this regularization mechanism, in opposition to the remaining. Given the dataset class imbalance, we opted to train the CNNs using a weighted binary cross-entropy (BCE). We penalize errors in the benign class with a weight $w_b = 0.02$, and $w_m = 0.98$ for the melanoma. The choice aims to reflect an approximation to the prior probability of each class. This strategy helped us not to converge to solutions where the classifier is biased towards the major class.

In a second phase, we were interested in testing the hypothesis of a CAE performing worse on the reconstruction task for the melanoma class. 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. The networks studied were two U-Net architectures. One with an encoder based on the Resnet-18 and the other on the Resnet-50. Our loss function was the mean squared error (MSE) between the input and the reconstructed images.

Finally, we closed the single image diagnosis with a one-class GAN. We tried two different architectures – deep convolutional generative adversarial network (DCGAN) and enhanced super-resolution generative adversarial network

[1] This analysis already excludes the 425 duplicated images listed in the challenge webpage [17].

(ESRGAN). As happened with the CAE, we wanted to exploit the dataset imbalance. In the GAN model, we went one step further by training the model with benign lesions, exclusively. Given a trained GAN, we tested the discriminator in a validation and a test set in which were also presented examples from the outlier class (melanoma).

2.3. Patient diagnosis

The next stage explains the framework oriented for the patient diagnostic. It uses intermediate results from the previous models, that were trained without information about the patient to whom each image belongs (single image diagnosis).

Having in mind building an automatic system to recognize UD lesions and make an intrapatient diagnostic with information from each patient context, we design a model to predict if a given patient has or not at least one melanoma among all its lesions. The neural networks optimized for the single image diagnosis can be used beyond the decision produced in the last layer since they capture patterns along the data forward pass and resume the input into feature maps. Specifically, we extract the features right after the CNNs global average pooling operation, which feed the fully connected layer, or from the discriminator's output map in the case of the ESRGAN. Those are the first summary of each image, typically vectors of dimension belonging to the range $\mathbb{R}^{361}$ to $\mathbb{R}^{2048}$, depending on the models. These feature vectors are still quite large and do not contain many crossed dependencies. For that reason, we propose a process to build patient embeddings from their various lesions feature vectors (Fig. 2). According to [3], when the dimensionality is quite large (as is our case), the classification models tend to perform worse, and so it is beneficial to reduce the dimensionality keeping relevant features. In our case, as we aim to detect outliers, we opted to resume our high dimension distance matrices using the five statistical operations presented in Fig. 2 since we think they may provide key information about the distance's distributions. We also investigated the possibility of incorporating in the embeddings the same five operations but performed on the CAE reconstruction errors, or even to combine embeddings obtained from several models.

Finally, once the embeddings are done, we can use them as inputs of a logistic regression model trained to predict the probabilities of that patient having at least one melanoma, or none at all. The logistic regression optimization is driven by a BCE loss. The reasonings for this loss function choice are the same as for the baseline image-wise diagnosis. However, in this case, we adapted the penalization weights used to $w_b = 0.2$ and $w_m = 0.8$ to reflect the prior probability of a patient having at least one melanoma among its lesions.

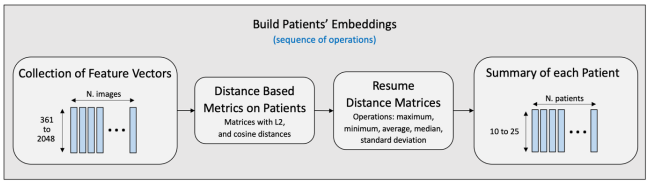

Fig. 2. From feature vectors to patient embeddings.

2.3. 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' results are concatenated to feed a fully connected layer that is responsible for weighing the contribution of each path in the final diagnosis. This pipeline is shown in Fig. 3. We apply a softmax activation to the last decision layer to get our final malignant scores for each input image. It is noteworthy that those scores already carry the intrapatient context.

After providing the big picture of the integrated diagnosis, we can go deeper with some more details concerning the training process. First, we load previously trained models when it is possible. This means that the upper path receives a trained baseline model or a GAN discriminator to do the image classification. Similarly, the lower path receives the logistic regression trained with the embeddings obtained from the model used in the upper path, or any combination.

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.2. After getting new embeddings, we train the logistic regression, exclusively, until notice convergence or reaching a maximum number of epochs. Then, in case the new patient diagnosis performs better than the previous, the general model is updated with this best new model. This way we distribute the training process to alternately optimize both paths.

Regarding the loss functions used, for the upper path and the combined output, we kept the loss function from the baseline diagnosis, and for the lower path we kept the loss used in the patient diagnosis. The only modification we 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,

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

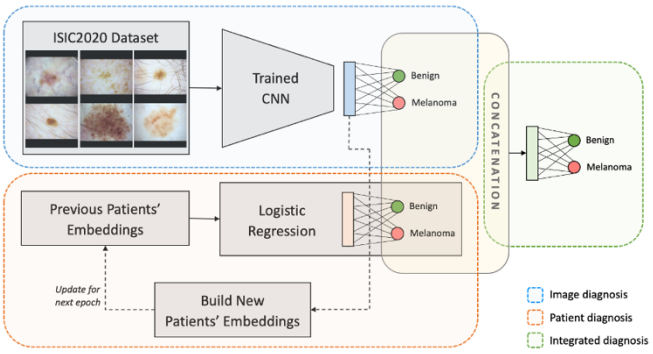

Fig. 3. Pipeline for the integrated diagnosis combining both the image and the patient diagnosis.

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.

3. EXPERIMENTAL SETUP

3.1. Online data augmentation

One way to improve the model generalization to new data is by having more diversity in the 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% of occurrence, followed by a random erasing. This last transformation replaces a rectangle of the image with black color pixels. The frequency of occurrence and the size and ratio of the rectangles was tuned as suggested in [11]. The only experiment that didn't use this transformation is the AE for lesion reconstruction since we did not want it to learn so much noise. However, it was possible to keep the random erasing if we had trained the AE for an inpainting task. Additionally, as the CNN initialization weights were imported from the ImageNet trained models, we normalized each image channel with the same mean and standard deviation of the ImageNet dataset as claimed in [12].

3.2. Dataset partition

We split the dataset to have a set of unseen data. It allows evaluate and compare the models across all the experiments, under the same set. The dataset 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 had approximately the same proportion of benign and melanoma lesions, with 98% and 2%, respectively, and in line with the whole dataset. Furthermore, the ISIC2020 challenge organization also provides a test set, but the ground truth labels are not publicly available. So, the test set can only be used to access the model's performance through the challenge submission site [13].

3.3. Computer hardware

To run our experiments we used a computer equipped with an Intel(R) Core(TM) i7-7700 CPU, 16GB RAM, and an NVIDIA GeForce GTX 1060 GPU with 6GB. Given this constraint, in some cases it was necessary to adapt the batch size. For example, the ESRGAN model was trained with a batch size equal to 2 due to memory capacity limitations. However, most of the time we used a batch size of 16.

4. RESULTS

4.1. Evaluation metrics

During the models' development and test, we need to be able to compare their performances. So, this section presents the mathematical expressions of the metrics we used for the image, the patient, and the integrated diagnosis.

As we are dealing with binary classification, benign versus melanoma lesions, we can define a confusion matrix with four entries – True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). Each entry tells the absolute frequency of the respective event after a model performance assessment. In an ideal scenario, where the model predicts all the examples correctly. However, most of the times, that is not the case, which leads to FP and FN diagnosis. We define True Benign Rate (TBR) and True Melanoma Rate (TMR) as

$$\begin{cases} \text{TBR} = \dfrac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{TMR} = \dfrac{\text{TP}}{\text{FN} + \text{TP}} \end{cases}. \tag{2}$$

Those metrics are meaningful for the imbalanced dataset because we can infer the ratio of benign and melanoma examples guessed correctly and not only global accuracy values. Furthermore, it is possible to combine these metrics and define the Balance Accuracy (BAC) like

$$\text{BAC} = \frac{1}{2}(\text{TBR} + \text{TMR}). \tag{3}$$

The BAC expresses a trade-off between the number of TN, TP, FN, and FP, crucial in medical applications. It should be noted that the TBR and the TMR are equally important in this definition, but it can be changed depending on the application

specifications by weighing each term properly. However, it only considers the predicted labels, not giving any hint concerning the output probabilities at the network's output. To bridge this issue, we consider the Area Under the Curve (AUC). This curve is the plot of the TBR against "1 – TMR" by changing the threshold value in the range [0; 1]. With this metric we can have an intuition about how far the predicted probability is from the ground truth class.

So far, we discussed the metrics used on the image-wise diagnosis. For the patient diagnosis, the criteria established requires guessing if the patient has at least one melanoma, or none. Then, the patient metrics concerning the TBR, TMR, and BAC follow the same reasoning as explained before.

4.2. Single image diagnosis

This section discusses the results obtained for single image diagnosis. The first subsection is dedicated to supervised learning (baseline models) and the following two subsections to unsupervised learning – a CAE and a GAN, respectively.

4.2.1. Baseline models

In this experiment we want to build a benchmark performance for the image diagnosis using CNNs without using the intrapatient context. From the set of CNNs studied EfficientNet-b2 is the one with the highest BAC (0.763) and AUC (0.870). Moreover, the TBR and TMR were equal to 0.918 and 0.607, respectively. Additionally, Resnet-18 is the network with the smallest gap between the TBR and TMR values (0.878, and 0.633, respectively). The BAC (0.756) and AUC (0.853) are just slightly below EfficientNet-b2.

After training the CNNs, we decided to proceed with a deeper analysis, starting by collecting the features from the global average pooling layer. Then we used them to do a Uniform Manifold Approximation and Projection (UMAP) into a two-dimensional plane. Finally, looking at the resulting distribution, we verified that our features were projected into some different clusters. This finding made us investigate what may be causing these separations and how it impacts the diagnosis. As a first guess we formulated the hypothesis of lesions from different body parts having unique characteristics which interfered with the networks' processing. However, the UMAP shows a fusion of several lesions' anatomic site in each cluster. So, we considered this first hypothesis invalid and formulated a new one. The second hypothesis is that the clusters may be related with the images' source. We already knew that images come from different hospitals, inclusively, their acquisition systems are different between them, apart from the New York and Brisbane hospitals. In fact, there is a clear separation across hospitals, except for the New York and Brisbane centers, as expected due to their similarity in the image acquisition conditions. This outcome let us validate our second hypothesis – the UMAP clusters are caused mainly due to the differences in the different image's sources. Fact that suggests that although we had used the Shade of Gray algorithm for image color constancy normalization, it was not sufficient to align these clusters and annihilate disparities among hospitals.

Given that our networks perceive the images' source in a different manner, we studied the performance of our best baseline model (EfficientNet-b2) across hospitals. These following results are detailed in Table 1.

4.2.2. CAE

We studied two U-Net architectures. One based on the Resnet-18, and the other on the Resnet-50. After training both networks, we inspected the MSE for the reconstruction task. Due to quite low MSE values, the input and the reconstructed images look the same, as if the reconstruction is perfect. We believe that this phenomenon is mainly due to the skip-connections between the encoder and the decoder. Those shortcuts made the training convergence much faster and stable. After plotting the MSE distributions through a boxplot for each network we noticed that the U-Net Resnet-18 has lower MSEs and more compact quantiles, when compared to the U-Net Resnet-50. The difference between the benign and melanoma quantiles is more evident in the first one. Additionally, we concluded that both U-Nets confirmed the hypothesis of performance decreasing for the reconstructing task of an unseen class (melanoma) at training time. Although, there might have been no advantage in increasing the Resnet depth, since the U-Net Resnet-50 does not have lower MSE, neither seems to improve the outlier class detection. Although the CAE information is not directly used for the image diagnosis, it will be useful later to help in the patients' characterization.

4.2.3. GAN

The first GAN tested was the DCGAN. However, even after trying some techniques to avoid the mode collapses, such as different learning rates for the discriminator and the generator, updating the generator more frequently than the discriminator, using batch normalization layers [14], applying data augmentation to the dataset, flip and smooth ground truth labels once in a while, and avoiding checkerboard patterns in the fake images by choosing a kernel size divisible by the stride in the transposed convolution layers, as suggested in [11], [15]. However, we could not reach a stable convergence between the generator and the discriminator. The main misfortune is related with the generator performance, the generated images are too noisy and blurred. There is also lack of diversity in the examples generated, which impacted the discriminator when we assessed the performance with the validation set, containing examples from both classes.

For that reasons we decided to try the ESRGAN approach described in [16] as the generator task seems to be easier when compared with the DCGAN. In fact, we noticed a huge improvement in the generated images. Fig. 4 shows a couple of examples.

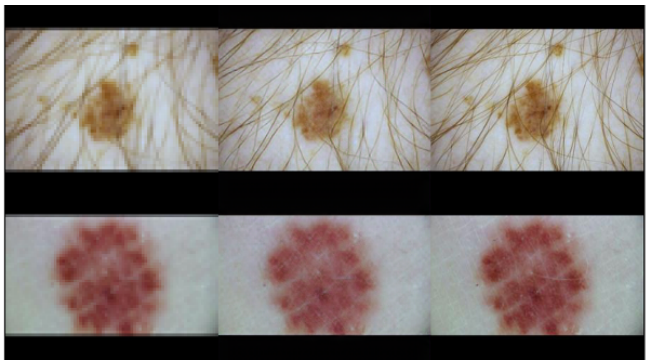

Fig. 4. Two examples where we can see from left to right the low-resolution version, the artificially generated image by the ESRGAN, and the original image. Besides the fake image appears quite realistic, we can notice colorization changes and blurred textures when comparing with the real image.

Although our main goal is not to enhance images, since we do adversarial training, the generator performance influences the discriminator. So, this result has good prospects. Effectively, the best discriminator performance in the complete validation set achieved a BAC equal to 0.826, with 0.711 and 0.940 for the TBR and TMR, respectively. The BAC is 6.3% higher than the best baseline model (EfficientNet-b2). The biggest change is in the TBR and TMR that dropped 20.7% and increased 33.3%, respectively. These results helped us to confirm the hypothesis of using a one-class GAN to detect the outlier (melanoma) class.

Similarly to the baseline diagnosis, collected features from the ESRGAN Discriminators' output map. It is still possible to see some clustering relative to the images' source hospital, although they are less evident than for the previous UMAPs. In addition, Table 1 shows the details concerning performance across the different sources.

4.3. Patient diagnosis

After studying several possible combinations to build our embeddings, our results showed that the best embeddings configuration joins information from the EfficientNet-b2 features and the U-Net Resnet-18 reconstruction errors. Adding the ESRGAN Discriminator features to the previous model seems to not improve the patient diagnosis, not only for the AUC but also for the TBR, and TMR. We expected to get a slight improvement in this case, but it might have occurred because the discriminator was trained exclusively with benign examples. In the end we got a patient diagnosis with the following performance metrics – BAC = 0.722, TBR = 0.706, and TMR = 0.739.

One interesting aspect to notice is the consequence of incorporating the U-Net Resnet-18 reconstruction errors. The CAE performs worser in the reconstruction task for the melanoma lesions, which leads to a shift in the distribution of the reconstruction errors when compared to the benign examples. As consequence the patient diagnosis improved when we take the CAE reconstruction errors into account. This fact helped us to validate the hypothesis of extracting meaningful images' characterizations with the CAE. Additionally, by inspecting the relative importance of each feature in the embeddings we saw that the CAE features represent about 95% of relative importance in the final decision.

Further, we present the patient diagnosis performance for models trained without context. Namely, we will discuss the results for the EfficientNet-b2 and the ESRGAN. To accomplish it we transform the results obtained by the previous networks for the image classification task, into a patient diagnosis that says if we guessed correctly that a given patient has a melanoma or not. Given this we obtained the following patient diagnosis for the baseline EfficientNet-b2 – BAC = 0.578, TBR = 0.554, and TMR = 0.602. The ESRGAN got BAC = 0.570, TBR = 0.173, and TMR = 0.966. Clearly, there is advantage in the patient diagnosis to optimize our model according to the patient context, as opposed to the single image diagnosis task. There is an increase of at least 0.11 for the BAC metric. Also, the patient diagnosis by inspecting the single image diagnosis output can easily bias towards one of the classes as happened for the ESRGAN. The low TBR and high TMR values are a consequence of the single image performance bias towards the melanoma class.

4.4. Integrated diagnosis

In our experiments we tested some combinations of previously trained blocks. Our choice was biased towards the models that had the best performances in their respective tasks. Particularly, we will present the results for the best combination. For the trained CNN we picked the EfficientNet-b2. For the logistic regression we opted to use the embeddings resulting from the EfficientNet-b2, and the U-Net Resnet-18 reconstruction errors.

By inspecting Table 1, we concluded that the integrated diagnosis outperforms the baseline diagnosis considering the BAC metric. The only exception is the ESRGAN Discriminator. Additionally, not only the performance by image increased, but also the performance by patient (obtained by transforming the results by image). This fact is again another clue that supports the hypothesis of existing advantage in including the patient context. It also shows that combining both diagnosis is constructive.

To translate this numerical overview into real examples, we present from Fig. 5 to Fig. 7 some examples of patients for whom integrating their context produced a better diagnosis outcome than the one achieved by the baseline diagnosis. For simplicity, we choose different scenarios where the integrated diagnosis helped. Meaning that corrections were made in different directions (FN and FP samples). The examples shown were evaluated on EfficientNet-b2 (baseline) versus

Table. 1. Performance per image and patient across hospitals. The best BACs between each of the networks are in bold.

	Network	Metric	Hospital					Global
			NYC	BCN	BNE	VIE	SYD	
Performance per Image	Baseline EfficientNet-b2	TBR	0.876	0.918	0.998	0.973	0.691	0.918
		TMR	0.821	0.371	0.000	0.250	0.676	0.607
		BAC	0.848	0.645	0.499	0.612	0.683	0.763
		AUC	0.930	0.716	0.498	0.820	0.701	0.870
	ESRGAN	TBR	0.810	0.563	0.908	0.337	0.851	0.711
		TMR	0.974	1.000	1.000	1.000	0.838	0.940
		BAC	**0.892**	**0.782**	**0.954**	**0.668**	**0.845**	**0.826**
		AUC	0.894	0.785	0.955	0.670	0.845	0.827
	Integrated Diagnosis (EfficientNet-b2 + U-Net Resnet-18)	TBR	0.873	0.896	0.999	0.966	0.542	0.902
		TMR	0.872	0.371	0.000	0.250	0.757	0.650
		BAC	0.872	0.634	0.499	0.608	0.650	0.776
		AUC	0.912	0.656	0.600	0.678	0.702	0.848
Performance per Patient	Baseline EfficientNet-b2	TBR	0.194	0.333	0.967	0.769	0.440	0.554
		TMR	0.793	0.400	0.333	0.250	0.739	0.807
		BAC	0.494	0.367	**0.650**	0.510	0.590	0.680
	ESRGAN	TBR	0.075	0.048	0.148	0.026	0.507	0.173
		TMR	1.000	1.000	1.000	1.000	0.870	0.966
		BAC	0.537	**0.524**	0.574	0.513	**0.688**	0.570
	Integrated Diagnosis (EfficientNet-b2 + U-Net Resnet-18)	TBR	0.224	0.286	0.967	0.756	0.413	0.545
		TMR	0.967	0.733	0.000	0.500	0.913	0.830
		BAC	**0.595**	0.510	0.484	**0.628**	0.663	**0.687**

on the EfficientNet-b2 plus the AE reconstruction errors (integrated diagnosis). Furthermore, to facilitate the figures interpretability, for each image we wrote the ground truth label over and painted a bounding box according to the predicted label – green for benign and red for melanoma.

To close this subsection, it is still missing a final consideration about the results across hospitals shown in Table 1. The integrated diagnosis performance per image is close to the EfficientNet-b2 (baseline), even the patterns described across hospitals. Also, in this assessment the best performance winner is the ESRGAN. However, the same cannot be said when we look at the patient performance. The major improvement claimed by the integrated diagnosis is specifically its advantage in the patient diagnosis performance without loss of performance in the single image diagnosis. On top of that, when we look to the performance per patient among the hospitals, the integrated diagnosis also wins the best BAC in two of them. And is the second best, just behind the ESRGAN, for two other hospitals. But the ESRGAN performance per patient is too unbalanced between classes, so we do not consider it a good predictor (for the patient diagnosis). To conclude, the performance per patient across hospitals also takes advantage with the introduction of the patient context.

4.5. Evaluation on the test set

Our goal in this section is to assess our model's performance on a test dataset. This will help us to discuss generalization. After introducing our model's predictions on the ISIC2020 challenge submission page [13], we got their respective scores on the private set. This process, and the fact that we do not have access to the test set ground truth labels make it impossible for us to show performances per patient. In this section, we can only report performances per image.

We tested three of our best models – the Baseline EfficientNet-b2, the ESRGAN Discriminator, and the Integrated Diagnosis based on the EfficientNet-b2 and the U-Net Resnet-18 reconstruction errors. Their performance on the test set is shown in Fig. 8. At a first glance, considering both the AUC and the BAC metrics, we notice that all models did perform worse in the test set than on the validation set. However, the discrepancy was significative mostly for the ESRGAN model (−26% in AUC). The major contribution was from a massive drop in the benign performance. The other two only dropped the AUC value up to 5%. Another common trend is the lower TMR results in the test set than in the validation set. In fact, the integrated diagnosis was the best model detecting the melanoma class, even though its TMR dropped from 0.667 to 0.590. In opposition, we showed that the TBR increased in the test set, approximately, up to 1% for the integrated and the baseline diagnosis, respectively. Both are close or above the 90% barrier of correctness in the benign predictions.

Overall, we consider our models did generalize well, except the ESRGAN. However, the performance discrepancy between both sets is not a complete surprise since we had already noticed instability in other application. Namely, when we used the previously trained ESRGAN discriminator in the integrated diagnosis. As we have seen, our networks perceived images' particularities across the different source hospitals. Given this, we conjectured that the test set may be a more challenging environment because it contains an additional source (Athens) that was not available at training time. In spite of that, we were able to successfully transfer the knowledge acquired at training time to a set of unseen data for at least two of our best models.

To close this analysis, we will compare our results with the ones obtained by the authors of [7] in the conditions described in section 1. In order to keep equity as much as possible we will use the results on the validation set as our reference, since despites in [7] they do a five-fold cross validation, their test dataset comes from the same data we have used in our work for training and validating models. Given this, the authors of [7] reported the following performance by image – AUC = 0.854, TBR = 0.770, and TMR = 0.766. Our baseline EfficientNet-b2, achieved a higher AUC, while the integrated diagnosis got an inline performance. Although in terms of performance balance per class, our models show a higher discrepancy. Regarding the performance by patient, we will compare their results with our results from the logistic regression models since the process for assessing performance is similar between both. In [7] they reported a performance by patient with BAC, TBR, and TMR all equal to 0.714. Our best logistic regression model got BAC = 0.722, TBR = 0.706, and TMR = 0.739, which is slightly better. This result might indicate that our approach to build patient embeddings based on distance-based metrics and statistical distribution metrics was similar effective as the strategy in [7] using a transformer encoder which is a much more complex pipeline.

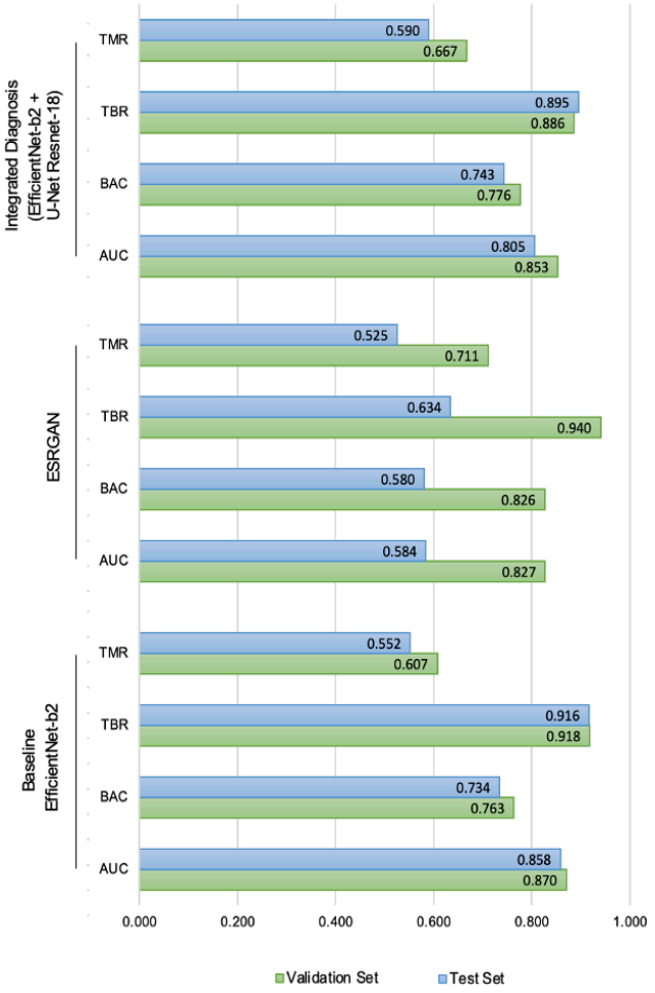

Fig. 8. Performance metrics (per image) considering both the validation and the test sets, among three diagnosis strategies.

5. CONCLUSION

5.1. Contributions

We addressed the problem of the melanoma skin cancer diagnosis. Our goal is to improve diagnosis by incorporating the intrapatient context and contribute to this rising research area. We used the medical UD concept, supported by clinicians, as the foundation for transferring it to both deep learning and anomaly detection fields.

Considering this, we tested an integrated diagnosis system that combines a diagnosis at the image and patient levels. To develop this final model, we aggregated supervised learning techniques, with the use of CNNs and Logistic Regression models, and self-supervised learning as is the case with the AEs and the GANs. This integrated system did successfully show a positive effect in combining several images from the same patient to improve the overall diagnosis. Accordingly, not only did the image diagnosis improve by over 2% in the BAC, but also the patient diagnosis by more than 10%, compared to baseline models which did not include the patient context. The best integrated diagnosis reported is composed of a trained EfficientNet-b2, and a logistic regression that receives patient embeddings built with deep features from the same network and information about the error in the reconstruction task using a previously trained CAE (U-Net Resnet-18).

As part of our research, we have also highlighted an interesting example of knowledge transfer using a GAN as a one-class classifier. Although it proved not to be so accurate with respect to the patient diagnosis, we have been able to improve the detection of melanoma. In the end this architecture got a higher AUC than any of the baseline diagnosis evaluated in the validation set. However, the one-class classifier GAN still shows lack of stability and generalization.

Finally, we have contributed with a couple of other relevant findings. After processing and projecting deep features extracted from some of our most robust networks, we noticed that the different lesions' anatomic site seems not to influence the diagnosis. The main source of biases is in fact discrepancies in the images caused by the different equipment at each hospital. This issue also influences the models' performance across the different hospitals. The integrated diagnosis was the best methodology found to minimize the performance per patient discrepancies across the different sources. Given this it constitutes another advantage in incorporating the patient context to help the diagnosis.

5.1. Future Work

Given some of these promising results, we believe there is still room to invest in different research directions. So, in this section we present some suggestions that may serve as future steps to continue this work.

i) Similarly, to the authors in [7], implement a mechanism to capture the patient context in an end-to-end fashion. We have developed a pipeline to build embeddings, but we cannot guarantee its optimality.
ii) Add to the integrated diagnosis a feature that points explanation to the output result. We think it would be helpful since it is difficult for a clinician to trust an automatic diagnosis without any other information besides an output label.
iii) Understand whether it is plausible to categorize patients based on their context (or phenotypic group). This is because there might not exist so many different groups, which opens up a chance to study some of them as a single entity.

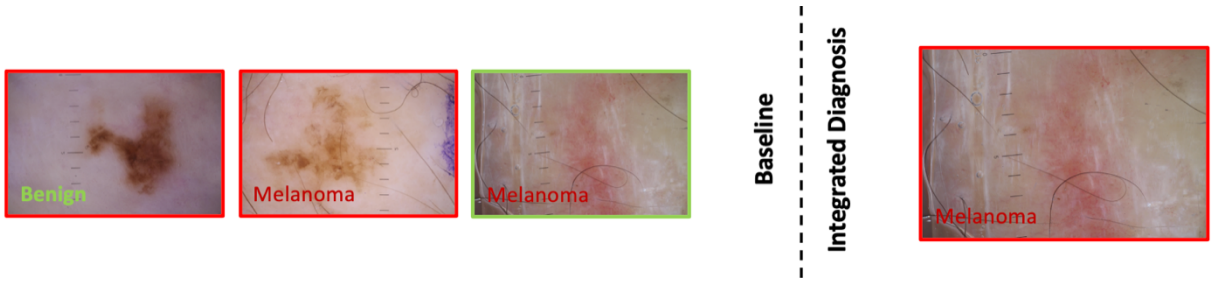

Fig. 5. Three lesions from a male patient who is 90 years old. On the left, the baseline diagnosis gave the right diagnosis for one of the melanomas, producing a FP and a FN. As a result of the integrated diagnosis, the FN diagnosis has been corrected, however, the label for the first image remains incorrect. We placed the ground truth labels over and painted a bounding box according to the predicted label – green for benign and red for melanomas.

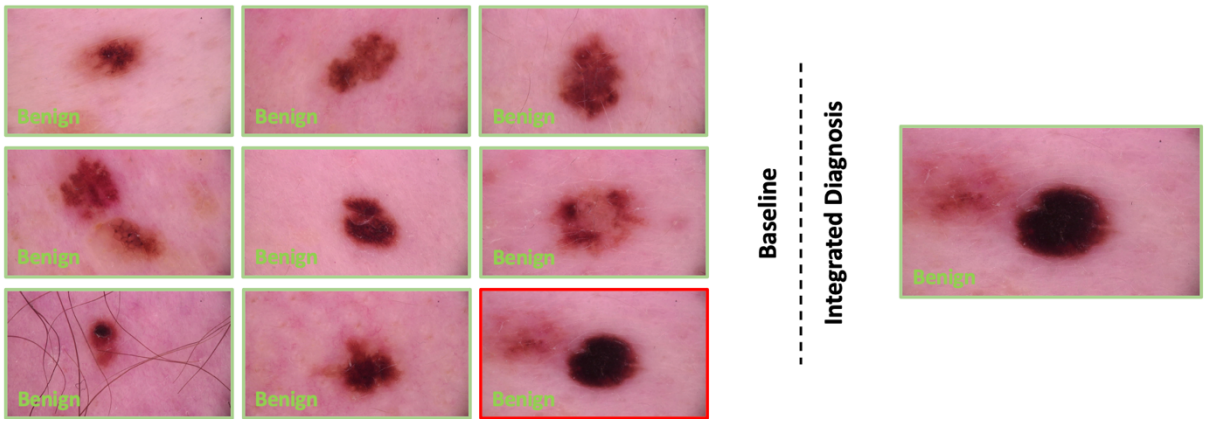

Fig. 6. Nine lesions from a male patient who is 55 years old. In this case, all lesions were benign, but the baseline diagnosis produced a FP. The mistake was corrected in the integrated diagnosis that ended up giving a full correct diagnosis for all the patient's lesions. We placed the ground truth labels over and painted a bounding box according to the predicted label – green for benign and red for melanomas.

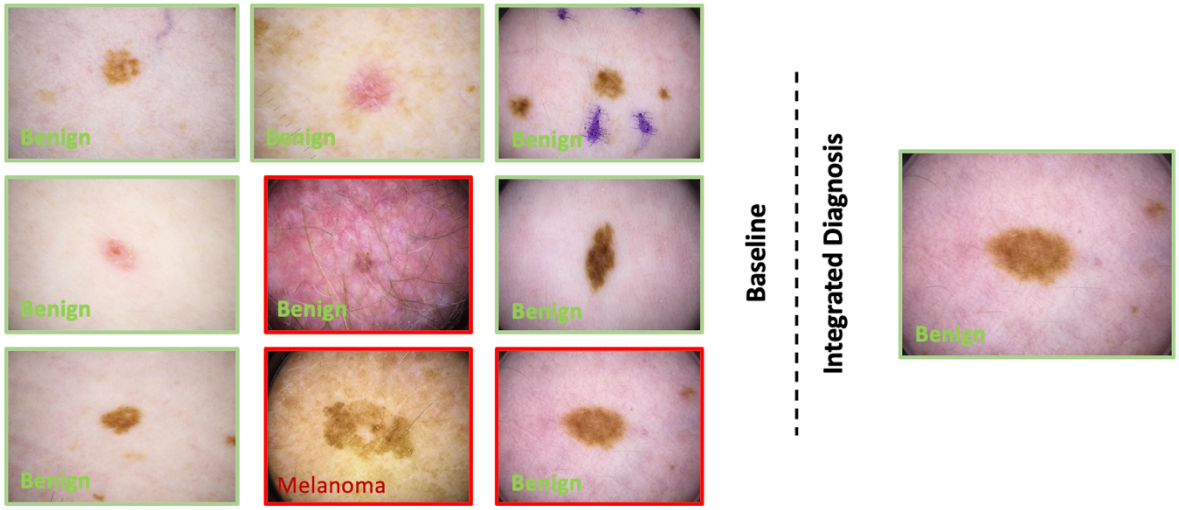

Fig. 7. Nine lesions from a woman who is 35 years old. Despite being able to detect only melanoma, the baseline diagnosis also produced two FP diagnoses. However, the patient context helped the integrated diagnosis discard one of the mistakes. It seems that FP not corrected may be a difficult assessment due to the significantly different colorization from the other lesions. We placed the ground truth labels over and painted a bounding box according to the predicted label – green for benign and red for melanoma.

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]. Available: https://challenge2018.isic-archive.com/leaderboards/

[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]. Available: http://stm.sciencemag.org/

[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] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.09088

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

[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 Std." https://github.com/developer0hye/PyTorch-ImageNet (accessed Mar. 15, 2022).

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

[14] "GAN training hacks." https://github.com/soumith/ganhacks (accessed May 21, 2022).

[15] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," Aug. 2017, [Online]. Available: http://arxiv.org/abs/1708.04896

[16] X. Wang *et al.*, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," Sep. 2018, [Online]. Available: http://arxiv.org/abs/1809.00219

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