Received: 30 July 2020 | Revised: 5 January 2021 | Accepted: 15 January 2021

IET Computer Vision

DOI: 10.1049/cvi2.12048

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

A deep analysis on high-resolution dermoscopic image classification

Federico Pollastri[1] | **Mario Parreño**[2] | **Juan Maroñas**[2] | **Federico Bolelli**[1] | **Roberto Paredes**[2] | **Daniel Ramos**[3] | **Costantino Grana**[1]

[1]Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia, Modena, Italy

[2]PRHLT Research Center, Universitat Politècnica de València, València, Spain

[3]Departament of Electronic and Communication Technology, Universidad Autónoma de Madrid, Madrid, Spain

Correspondence

Costantino Grana, Dipartimento di Ingegneria 'Enzo Ferrari', Università degli Studi di Modena e Reggio Emilia, Via Vivarelli 10, Modena, 41125, Italy.
Email: costantino.grana@unimore.it

Funding information

Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER), Grant/Award Number: IDIFEDER/2018/025

Abstract

Convolutional neural networks (CNNs) have been broadly employed in dermoscopic image analysis, mainly as a result of the large amount of data gathered by the International Skin Imaging Collaboration (ISIC). As in many other medical imaging domains, state-of-the-art methods take advantage of architectures developed for other tasks, frequently assuming full transferability between enormous sets of natural images (e.g. ImageNet) and dermoscopic images, which is not always the case. A comprehensive analysis on the effectiveness of state-of-the-art deep learning techniques when applied to dermoscopic image analysis is provided. To achieve this goal, the authors consider several CNNs architectures and analyse how their performance is affected by the size of the network, image resolution, data augmentation process, amount of available data, and model calibration. Moreover, taking advantage of the analysis performed, a novel ensemble method to further increase the classification accuracy is designed. The proposed solution achieved the third best result in the 2019 official ISIC challenge, with an accuracy of 0.593.

1 | INTRODUCTION

Skin cancer is a major public health issue, being the most common forms of human cancer worldwide [1]. Malignant melanoma is less common than basal and squamous cell carcinoma (it accounts for only about 3%–4% of all skin cancers), but it is responsible for most of the deaths [1]. Despite all the advances in skin cancer treatments, early detection remains a key factor in preventing their progression to advanced stages and thus lowering the mortality rate [2].

To perform a fast diagnosis, many dermatologists rely on dermoscopy, which is a form of in vivo skin surface microscopy performed using high-quality magnifying lenses and a powerful light source to mitigate the surface reflection of the skin, in order to enhance the visibility of the pigmentation of the lesion (Figures 1 and 2). This imaging technique has increased the diagnosis accuracy, sensitivity, and specificity with respect to the naked eye examination, mitigating the need of surgical intervention for the unnecessary removal of benign lesions. However, to diagnose skin cancer through this kind of non-invasive imaging approaches, a thorough image analysis must be performed by expert clinicians. This is why many efforts have been carried out in recent years towards the creation of tools to assist physicians in the analysis of dermoscopic images. Deep learning in particular, due to its outstanding results in many areas such as speech recognition [3], image understanding [4] and image classification [5, 6], has become the main option for analysing medical images.

A bigger neural network size has always been considered a synonym for better accuracy, since a larger amount of parameters and layers means a greater capability to learn important filters, and therefore it is possible to capture meaningful features within an image as long as its resolution is high enough. However, deeper networks are more prone to overfitting and harder to regularise during training, in addition to requiring a considerable amount of time to be trained [7]. This often produces inefficient architectures, which do not improve results yielded by faster, shallower networks. Tan et al.

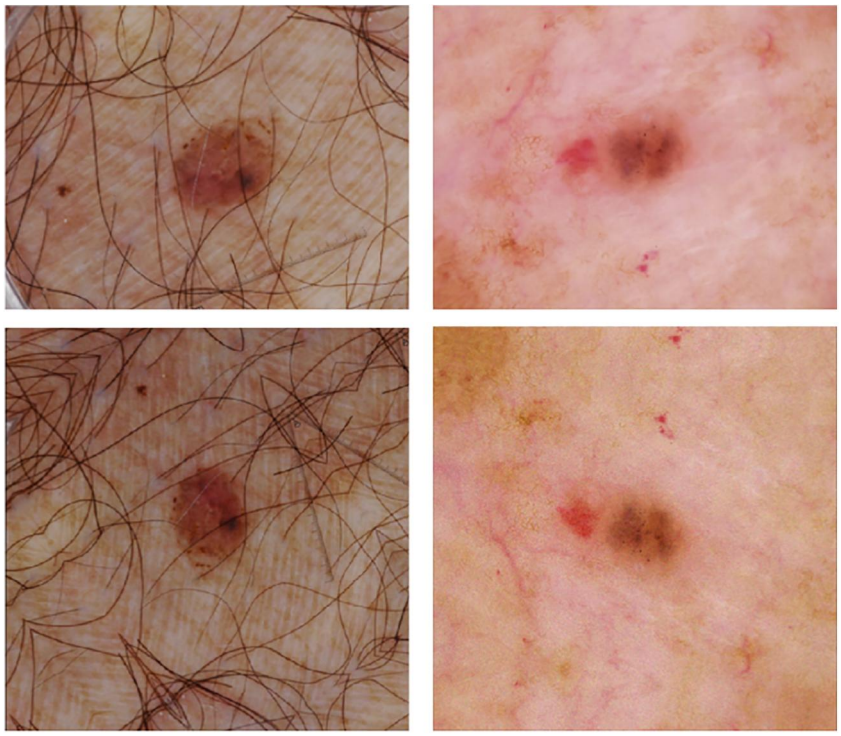

FIGURE 1 Random samples of dermoscopic images (top), coupled with the results of random data augmentation (bottom) performed by flipping and rotating the images, applying Gaussian filters, adding noise with a Poisson distribution, and manipulating hue, and saturation

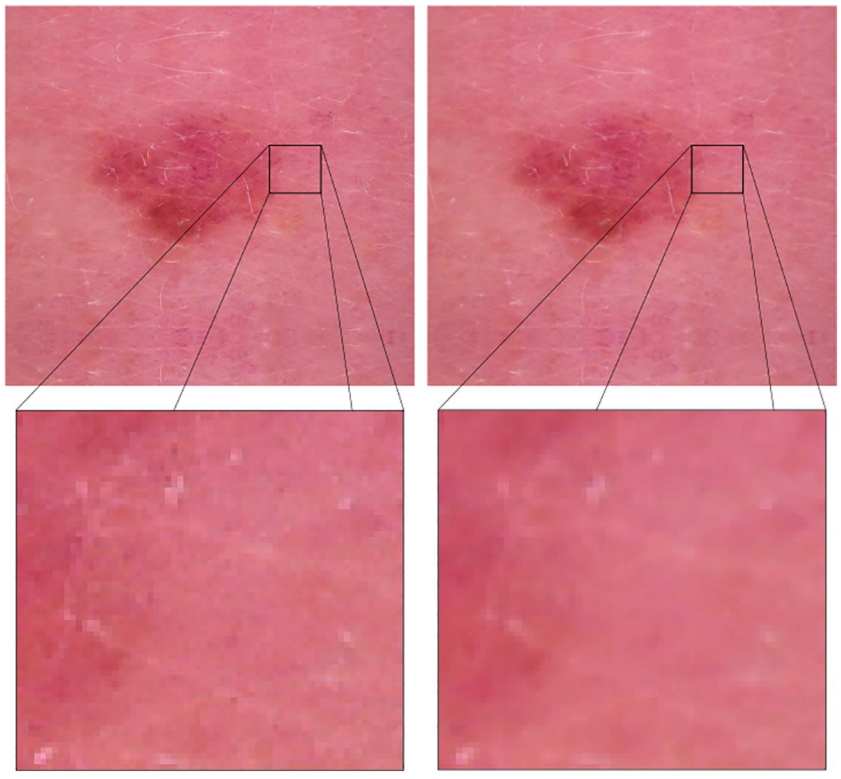

FIGURE 2 On the left, the original 512 × 512 input image and an enlarged detail. On the right, the same image after a blurring filter is applied, with the same enlarged patch. The original image presents a noisy checkerboard effect, introduced by a sharpening filter. This effect is greatly reduced in the picture on the right

[6] proposed an efficient way to scale up convolutional neural networks' (CNNs) width, depth, and image resolution by performing exhaustive experiments on several datasets of natural images.

However, medical images present several dissimilarities from natural ones, and thus require an individual analysis [8]. As a matter of fact, in dermoscopic images, the difference between background (human skin) and foreground (skin lesion) can be less visible than in most other scenarios. Indeed, sharpening filters are broadly used in this field to enhance lesion borders. Moreover, dermoscopic images present numerous artefacts such as black round borders, pen drawings,

rulers, and hair, which must be ignored when seeking meaningful patterns within an image.

This particular domain of medical imaging is characterised by a very high resolution, which must be taken into account when aiming for good classification accuracy. Additionally, a huge quantity of data augmentation (DA) strategies can be performed without altering the nature of the skin lesion, and this can be exploited during inference by merging the outputs of models that are robust against different combinations of simple transformations.

Finally, it is crucial to make use of calibrated models for a correct behaviour in critical decision scenarios, such as medical diagnosis, in which the ultimate goal is not substituting an expert practitioner but providing a reliable measure of our degree of uncertainty to assist the final decision.

The main contributions of this work can thus be summarised as follows:

- We perform a thorough investigation about the performance of state-of-the-art architectures for natural images classification when applied to dermoscopic image analysis.
- A comprehensive discussion on how the major hyperparameters (the size of the network, image resolution, DA process, amount of available data, and model calibration) affect neural network capabilities is provided.
- We explore and motivate the use of model calibration to improve the overall accuracy of a deep learning architecture in skin lesion analysis.
- We design a novel ensemble method for dermoscopic image classification, which yields a balanced accuracy of 0.593 on the official 2019 ISIC challenge, achieving the third best result.
- The first classified algorithm of the official 2019 ISIC challenge is compared with the proposed method. Experimental results show that our approach outperforms the winners of the challenge when the two algorithms are trained and tested using the same data.

The rest of this article is organized as follows. In Section 2, a detailed description about relevant literature is presented. Section 3 introduces the 2019 ISIC dataset and the proposed preprocessing pipeline. The designed ensemble architecture is presented and motivated in Section 4, and evaluated in Section 5 through an in-depth analysis. Finally, in Section 6 conclusions are drawn.

2 | RELATED WORK

2.1 | Dermoscopic diagnosis

Skin cancer is the most common cancer all around the globe, with melanoma being the deadliest form [1]. Dermoscopy is a skin imaging modality that allows for a better skin cancer diagnosis, with respect to unaided visual inspection. However, clinicians must receive adequate training for these improvements to be achieved. To address this, multiple organisations such as the International Skin Imaging Collaboration (ISIC) [9, 10], have released dermoscopic images datasets specifically designed for deep learning, labelled with different skin lesions categories.

Since 2016, ISIC started hosting challenges and workshops, gathering new images and annotations every year and focussing on different tasks, ranging from lesion segmentation and lesion attribute detection to disease classification. An in-depth description of the 2019 version of the dataset, which is employed to perform the experiments described in this article, is provided in Section 3.

2.2 | Classification CNNs

CNNs have become the dominant machine learning approach, and the *scaling up* strategy has been widely used to achieve better accuracy results. As an example, ResNet [11] can be scaled up from ResNet-18 to ResNet-200 just by adding more layers. However, this technique leads to the notorious problem of vanishing/exploding gradients [12], which hampers the convergence of the architectures. This problem has been managed with different approaches, such as intermediate normalisation layers [13] or normalised initialisation [14, 15], resulting in great accuracy improvements over the years. In 2016, Xie et al. proposed ResNeXt [16], which introduces the concept of *cardinality*, by slightly changing the residual block structure proposed with ResNet. In the same year, DenseNet [17] proposed a new architecture which increases the number of connections of each layer, alleviating the vanishing-gradient problem. SEResNeXt [18], introduced in 2017, provides significant performance improvements with respect to existing state-of-the-art CNNs, by means of 'Squeeze-and-Excitation' (SE) blocks, that adaptively recalibrate channel-wise feature responses. Finally in 2019 EfficientNet [6], provided a new scaling up method that uniformly increases the dimensions of depth, width, and resolution, achieving state-of-the-art accuracy on ImageNet [19].

2.3 | High accuracy

Transfer learning has become a de-facto method to enhance the training process of deep learning models. It is a vital tool to be exploited when the data in the target domain is not abundant. Natural image datasets such as ImageNet are usually chosen to pre-train neural networks, due to the vast amount of labelled data. However, the performance can worsen when the source and target domains present several differences [20, 21]. Fortunately, it is frequently possible to take advantage of these pre-trained features and increase the network accuracy thanks to the transferability of simpler filters, such as those that address colour, size, or edges [2]. Furthermore, the performance of a CNN can be boosted through an extensive investigation of the hyperparameterisation and using ensembles of several models [22, 23]. The latter approach is stated to produce the best accuracy results [24], despite a heavy computational cost in terms of resources, training time, and inference time. However, in the skin lesion analysis domain, reliable results must be preferred to low inference time.

2.4 | Deep learning in medical imaging

Deep learning methods have been employed in several medical fields, such as renal biopsy [25], image retrieval [26], and the detection of multiple forms of cancer [27].

As a matter of fact, deep learning-based methods have also been proposed to tackle dermoscopic image analysis [28, 29]. In 2019, Wang et al. introduced an enhanced high-level parsing (EHP) module to generate meaningful feature representation for skin lesion [30]. The following year, the same main author investigated the complex correlation between skin lesions and their informative context by placing a bi-directional dermoscopic feature learning module on the top of a CNN network [31]. Furthermore, skin lesion boundaries segmentation CNNs can be adopted to improve classification accuracy, by removing non-prominent features from dermoscopic images [23]. This technique allows multiple lesions within a single image to be correctly extracted and classified.

Finally, Gessert et al. provided a description of the best performing approach for the 2019 ISIC challenge [32]. The authors employed several versions of EfficientNet and made use of an ensemble method to obtain the final prediction. Moreover, a pre-processing technique is applied to remove black corners from dermoscopic images, and two different input strategies are used: same-sized cropping and random-resize cropping. Unfortunately, an additional private dataset is exploited during the training process, making it impossible to reproduce the experiments reported in the article.

3 | DERMOSCOPIC IMAGES

Since 2016, the International Skin Imaging Collaboration (ISIC) has begun to aggregate a large-scale, publicly available dataset of dermoscopic skin lesions images (Figure 3) and is hosting multiple challenges and workshops [9]. The availability of this substantial amount of dermoscopic images allowed is to significantly improve the performance of machine learning algorithms. This dataset, also known as the ISIC archive, is designed both for clinical training and to support research towards automated skin cancer analysis.

The 2019 version of the ISIC archive contains a total amount of 25,331 dermoscopy labelled images, belonging to eight different classes (i.e. types of skin lesion) [33]. Images have been collected in several years, from different centres, and using multiple devices. For these reasons, their resolution ranges from 450 × 600 to 1024 × 1024 pixels.

The available data is heavily imbalanced, as samples are distributed among classes as follows:

1. Melanoma (MEL) – 17.8%
2. Melanocytic Nevus (NV) – 50.8%
3. Basal Cell Carcinoma (BCC) – 13%
4. Actinic Keratosis (AK) – 3%
5. Benign Keratosis (BKL) – 10%
6. Dermatofibroma (DF) – 0.9%
7. Vascular Lesion (VASC) – 1%
8. Squamous Cell Carcinoma (SCC) – 2.4%

However, the official 2019 test dataset counts an additional class, which is not available in the training partition of the data. This class, named *none of the others*, contains dermoscopic images of different natures that do not belong to any of the other eight classes. To correctly evaluate such a heavily imbalanced task, the 2019 official challenge judges the participants by means of the Balanced Accuracy metric, which is computed as the average sensitivity among classes. This metric gives the same importance to each class, regardless of how

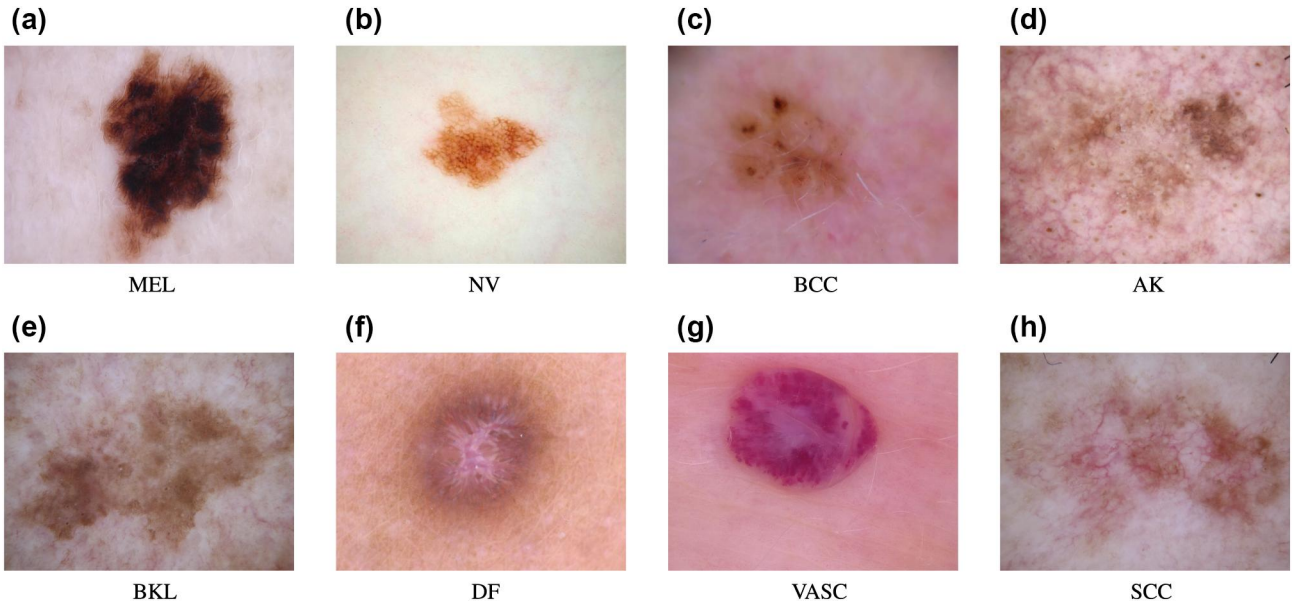

FIGURE 3 Samples of the 2019 ISIC dataset. (a) Melanoma – MEL, (b) Melanocytic Nevus – NV, (c) Basal Cell Carcinoma – BCC, (d) Actinic Keratosis – AK, (e) Benign Keratosis – BKL, (f) Dermatofibroma – DF, (g) Vascular Lesion – VASC, (h) Squamous Cell Carcinoma – SCC

much it is represented in the test set. Additional metrics such as the Area Under the ROC Curve (AUC) are presented in the official leaderboard, but not employed in the final scoring.

3.1 | Preprocessing

Dermoscopic images present several dissimilarities from natural images such as a very high resolution, low colour variability within an image, and many unnatural artefacts like pen marks or black corners introduced by acquisition devices. This is mainly because the subject of these images is human skin, and because of the particular acquisition technique that aims to manipulate how the light hits the epidermis. The discrepancy can be noticed by just computing the dataset statistics and using Gaussian distributions as an approximation of the real pixels values' distribution (Figure 4). It is thus crucial to carefully choose preprocessing steps and DA strategies, instead of reusing procedures developed for natural images. Hence, a specific dataset mean and standard deviation must be computed and exploited for input normalisation, which is an essential step to obtain an efficient training process. To ensure good results, the same values must be used to perform input normalisation during inference.

Four main characteristics are widely recognised as primary attributes for the detection of melanocytic lesions, which are asymmetry, border irregularity, colour variegation, and a diameter greater than 6 mm [34]. The analysis of these four features is also known as the ABCD rule, and it represents the basic guideline to preserve semantic information within dermoscopic images. Since image sizes are not constant, the first step is to obtain a dataset of squared images by replicating the border of rectangular pictures along the shorter side, to not change the shape nor the size of skin lesions, which are key factors for the diagnosis. The next step concerns DA, an important and well-known operation that can be performed during training to improve the effectiveness of neural networks [35]. This process consists in generating new data items by applying very simple transformations to existing training samples, without changing their semantic content. This technique aims to improve the robustness of an algorithm against used transformations and thus boosts the final accuracy [36, 37]. With respect to natural images, dermoscopic ones can benefit from a larger amount of DA steps, considering that transformations like vertical flips or rotations do not alter their semantic content. Furthermore, dermoscopic images are often refined by means of sharpening filters to emphasize lesion borders and make images easier to inspect for expert dermatologists. This processing technique can however lead neural networks to erroneously focus on low-level features that are trivial. To avoid this drawback, images can be randomly blurred through Gaussian kernels, teaching CNNs to ignore differences caused by the use of diverse sharpening filters. Figure 2 shows an example image before and after the employment of the blurring filter; the enlarged version of the images exposes that low-level artefacts introduced by the sharpening filters get mitigated thanks to the Gaussian filters. Blurring images with random values allow trained CNNs to increase their steadiness against a wide spectrum of sharpening filters that are used during acquisition.

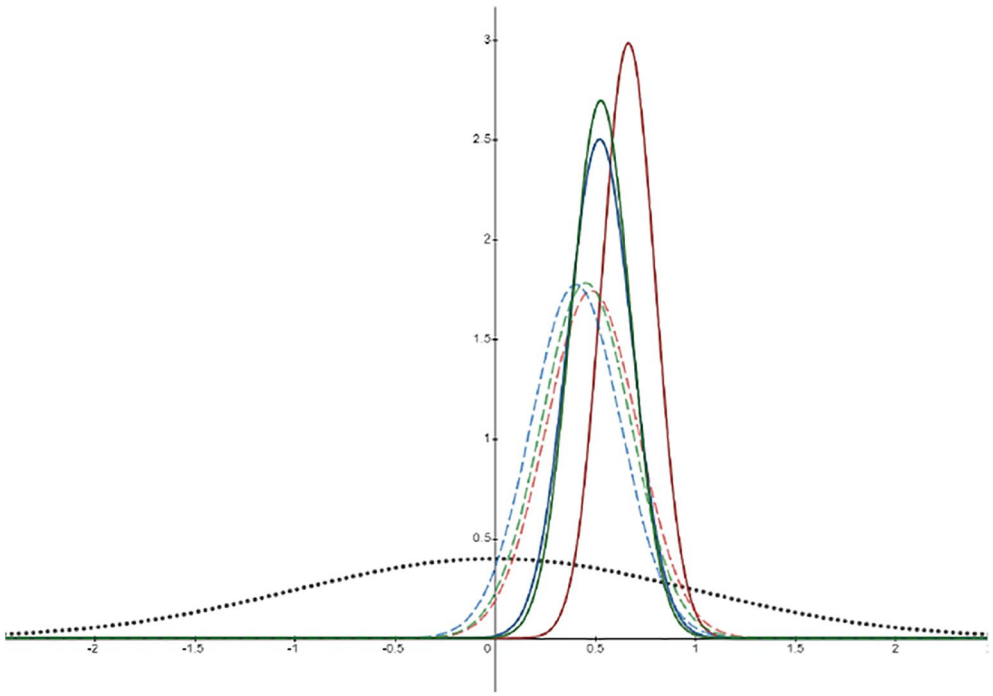

FIGURE 4 Approximation of the real pixel values distribution through Gaussian distributions. Dashed, lighter lines represent the distribution of channels R, G, and B of the ImageNet Dataset, whereas solid, darker lines represent the ISIC dataset. The grey dotted line is the Gaussian distribution with mean 0 and standard deviation 1, which we aim to obtain after input normalisation

TABLE 1 Different data augmentation configurations. When not specified, configuration 0 is employed

Configuration	Flips	Rotating	Gaussian Filter	Cutout	Add Noise	Hue and Saturation	Contrast
0	✓	✓	✓	✓	✓	✓	✗
1	✓	✗	✗	✗	✗	✗	✗
2	✓	✓	✓	✓	✗	✗	✗
3	✓	✓	✓	✓	✓	✓	✓

TABLE 2 Training times (expressed in minutes), balanced accuracy, and area under the ROC curve (AUC) of several neural networks architectures with different image sizes

	512 × 512			256 × 256			128 × 128		
Net	Training Time	Balanced Accuracy	AUC	Training Time	Balanced Accuracy	AUC	Training Time	Balanced Accuracy	AUC
DenseNet-201	5700	0.862	0.980	1870	0.820	0.975	1550	0.752	0.950
DenseNet-121	3420	0.834	0.972	1619	0.809	0.966	1505	0.739	0.942
SEResNeXt-101	9804	0.857	0.981	2782	0.831	0.978	1687	0.745	0.960
SEResNeXt-50	5928	0.867	0.982	2006	0.818	0.975	1642	0.761	0.958
ResNet-18	2052	0.806	0.975	1573	0.789	0.968	1482	0.708	0.937
ResNet-50	3192	0.841	0.977	1641	0.796	0.965	1550	0.692	0.930
ResNet-152	7752	0.861	0.978	2280	0.802	0.972	1687	0.723	0.946

Abbreviation: AUC, area under the roc curve.

As previously mentioned, randomly flipping and rotating images during training never change the semantic content, and always yield good results. Moreover, the manipulation of contrast, hue, and saturation are very common techniques for augmenting these kind of images. This is because of the assumption that different acquisition devices can alter the representation of similar colours, in addition to the effect that different light settings and natural skin tones can have on images captured with the same camera. Even though we find accuracy gains yielded by these DA strategies to often be minor, they can all be exploited to increase the robustness of the inference method described in the following Section.

Finally, the network training regularization can be improved through the CutOut strategy [38], and by adding Poisson distributed random noise to the input image. Figure 1 presents two samples of dermoscopic images, and their appearance after applying the random DA described in the first row of Table 1, the effects of the cutout method are not displayed as it is applied after pixel normalization.

4 | HIGH ACCURACY THROUGH GROWTH

To improve classification accuracy, CNNs can be scaled up in multiple ways. The most common approaches can be summarised as increasing either the number or the size of convolutional filters within a model. In addition to the growth of time required to complete the training process, scaled up networks necessitate a more careful regularisation, and are more prone to overfitting [7]. Both of these problems are especially relevant when the volume of available training data is not large enough. On the other hand, increasing the input image resolution also boosts CNN performance regardless of the network architecture employed, at the cost of incrementing the required training time. Adopting a larger input size makes it possible to fully exploit the high resolution that characterises dermoscopic images. As a matter of fact, CNNs can be effectively fine-tuned by using images of a different resolution than the one used during the pre-training process, Tables 2 and 3 show how increasing resolution is the most effective way to boost the skin lesion classification accuracy for every tested architecture (EfficientNet [6], ResNet [11], DenseNet [17], and SEResNeXt [18]). As depicted in Figure 5, growing the input size from 256 to 512 yields improvements showcased by balanced accuracy boosts that range from 1.7% to 5.9%, whereas growing the network size by increasing the number of parameters produce smaller boosts and, in some cases, even drops in accuracy.

4.1 | Probabilistic model

In medical contexts, it is crucial to provide both good discriminative power and reliable confidence. Therefore, in

Net	Image size	Training time	Balanced accuracy	AUC
EfficientNet-b0	224	1573	0.758	0.961
EfficientNet-b1	240	1642	0.796	0.967
EfficientNet-b2	260	1824	0.818	0.971
EfficientNet-b3	300	2280	0.830	0.974
EfficientNet-b4	380	4332	0.836	0.978
EfficientNet-b5	456	9348	0.831	0.975
EfficientNet-b6	528	17,328	0.829	0.968
EfficientNet-b7	600	29,640	0.827	0.967

Abbreviation: AUC, area under the roc curve.

TABLE 3 Scaled EfficientNet performance results. Training times are expressed in minutes

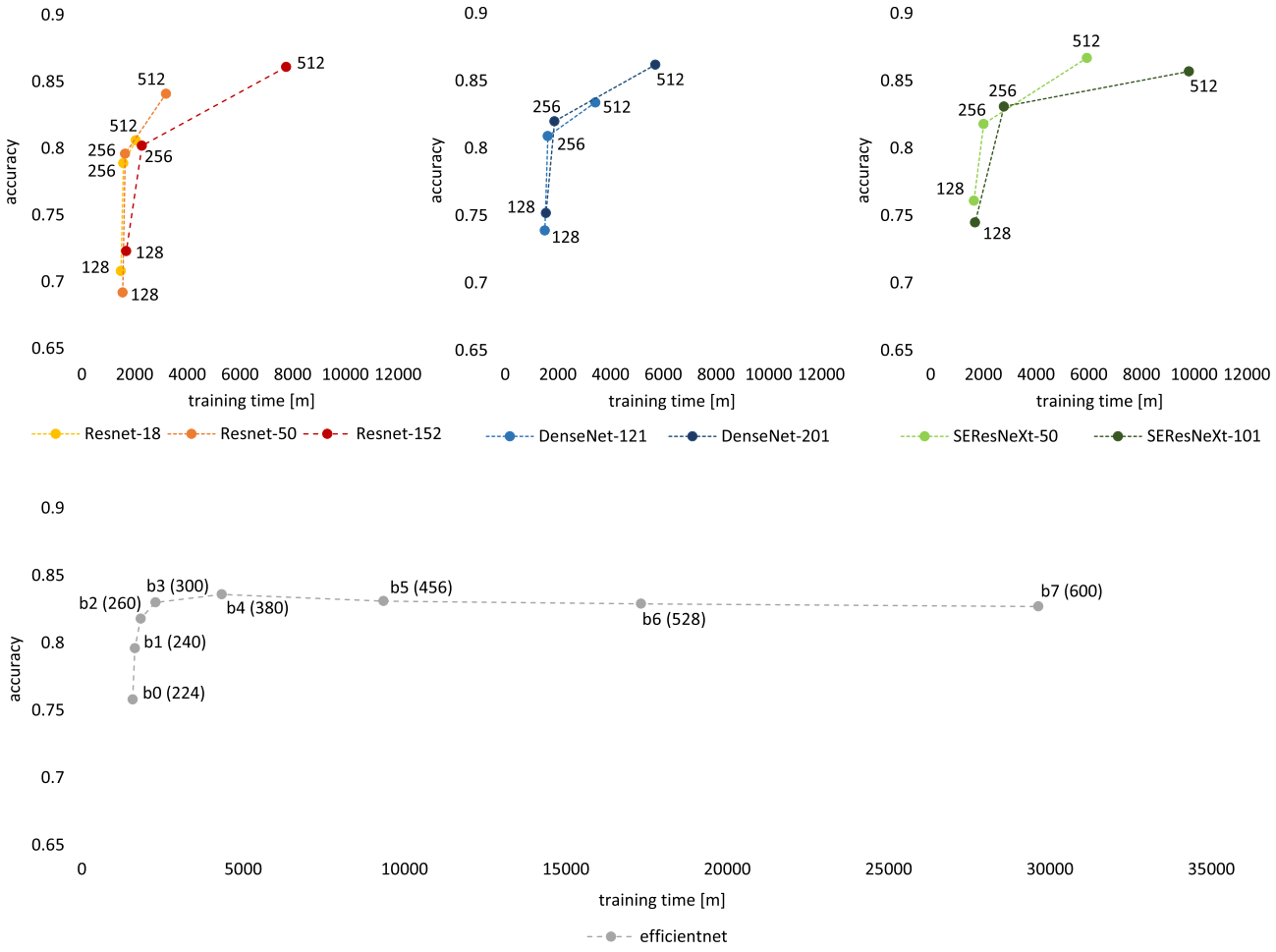

FIGURE 5 Correlation between **balanced accuracy** (*y* axis) and **training times** in minutes (*x* axis) for different network architectures. Within charts, each point represents one trained model, characterized by the network architecture and the input image size. EfficientNet CNNs are trained with their conventional input image resolution, whereas every other network is trained with three different image sizes: 512 × 512, 256 × 256 and 128 × 128. Lower training times are required for the smallest resolution, and vice versa

this section we describe our probabilistic model, which is divided in how we assign probabilities, and how we combine them.

Given a set of independent and identically distributed (i.i.d.) labelled pairs of samples $\mathcal{O} = \{x_i, t_i\}_{i=1}^{N}$ made up from images x_i with their corresponding categorical class labels t_i, we estimate the joint distribution of a set of M models, also known as the model ensemble, parameterised by $\Theta = \{\theta_m\}_{m=1}^{M}$.

Given the model parameters θ, assuming independence between the models and a non-informative prior over θ, the joint distribution factorises as:

$$p(t^1, t^2, \ldots, t^M, \Theta|\mathcal{O}) = \prod_m p(t^m|\mathcal{O}, \theta_m) \cdot p(\theta_m) \quad (1)$$

Learning involves maximising this joint model, which under our assumptions is the same as learning each of the conditional distributions $p(t^m|\mathcal{O}, \theta_m)$ separately, by optimising the cross entropy loss without any form of regularisation. The proposed method maps each of these conditional distributions with different network architectures and DA techniques. The optimal number of networks and preferred model architectures are chosen using a validation set, because using the evidence framework [39] for automatic Occam's razor is intractable, and approximations are a computational burden. The final probabilistic vector assigned to a given test sample x^* is obtained by computing the posterior distribution of the label, given the models $p(t^*|t^1, t^2, \ldots t^M, x^*)$. This posterior can be computed in several ways: with standard model average [40], with a learnt combination, with boosting techniques [41] or Bayesian classifiers [42], that is, substituting the weights by the posterior distribution of the models given the data.

Considering standard model average, for a test sample x^* we assign the label t^* with confidence p^* as follows:

$$\begin{aligned} p(t^*|x^*, t^1, t^2, \ldots t^M) &= \frac{1}{M}\sum_m p(t^m|x^*, \widehat{\theta}_m) \\ \widehat{t}_i &= \operatorname{argmax} p(t^*|x^*, t^1, t^2, \ldots t^M) \\ \widehat{p}_i &= p(t^* = \widehat{t}_i|x^*, t^1, t^2, \ldots t^M) \end{aligned} \quad (2)$$

Moreover, we augment this posterior probability with a small set of models for which we perform DA at inference time, that is we combine the predictions of modified versions of a given test sample x^* through image transformations, as described in Section 4.2.

However, one of the consequences of doing Maximum A posterior Probability (MAP) estimation is that, when dealing with unbalanced datasets, many local optima tend to ignore the unrepresented classes. A possible solution is to compute the predictive distribution under a Bayesian paradigm, but this is again impractical for our purposes. The solution we adopt is to turn the discriminative classifier into a generative one, and subtract the prior information over the classes:

$$p(t_1|x, \theta) \propto p(x|t_1, \theta) \cdot p(t_1|\theta) \quad (3)$$

Then, by scaling the posterior probability by the inverse of the prior information $1/p(t_1|\theta)$ we force our network to learn $p(x|t_1, \theta)$. This can be viewed as learning the posterior probability $p(t|x)$ assuming equal prior distributions $p(t)$, thus, training the model to not discard the unrepresented class. The model evidence $p(x)$ plays no role in this re-scaling as it is common to all the classes, thus it can be absorbed in the learning rate. In practice, this can be efficiently done by a weighted cross entropy loss [43].

In the medical diagnosis field, the goal should not be to take an action on behalf of an expert practitioner, but rather to assist his/her choice [44]. We thus end up discussing model calibration, mandatory for optimal decision. In such scenario, the decision made by an expert practitioner is differently influenced if we provide a confidence of 0.9 over a confidence of 0.4. This means that the provided information will only be useful if it is reliable, and this is achieved by a proper calibration of the probabilistic model. For a wider description of model calibration in a classification scenario see [45] and references therein.

The intuitive motivation of having a calibrated model can be seen from a more theoretical perspective. Taking into account that our aim is to combine expert knowledge and probabilistic information in an optimal way, we can formalise the problem using the Bayes decision rule, where the selected action α_i is the one that minimises Bayes risk:

$$\begin{aligned} \mathrm{R}(\alpha_i|x) &= \sum_j \lambda_{ij} p(t_j|x) \\ \alpha_i &= \operatorname*{argmax}_i \mathrm{R}(\alpha_i|x) \end{aligned} \quad (4)$$

$R(\alpha_i|x)$ denotes the Bayes risk and λ_{ij} is the loss incurred in deciding class i when the true value is j. Note that for the particular case in which $\lambda_{ii} = 1$ and $\lambda_{ij} = 0$ this rule ends up being the maximum a posterior decision rule. Furthermore, expert knowledge can be incorporated in these coefficients.

It is well known that this rule guarantees optimal decision if the data reflects the distribution $p(t|x)$ [41]. In practice, these distributions are substituted with our model $p(t|x, \theta)$. This means that the lower the gap between the model and the data distribution, the closer we are to an optimal decision. In general, this is achieved by models both able to separate between classes (a property known as discrimination or refinement [46]), and able to assign correct probabilities based on how the data is distributed (calibration).

Motivated by the properties of calibrated models, we propose to ensemble neural networks that have been previously calibrated. Considering that ensembles aim to combine probabilistic information, one should expect that the combination of more reliable classifiers, as those with a proper calibration, should provide more reliable final posterior probability (Equation 4). In fact, merging multiple classifiers usually boost the accuracy, as the combination of different local minima is a better representation of the data distribution [47, 48]. Following this observation, it has been proved in [40] that the average combination of classifiers tends to also calibrate the final predictions. We extend this consideration using an ensemble of calibrated models.

4.2 | Averaging calibrated probabilities

We observe that DA can be employed also during inference to provide calibrated probabilities. As a matter of fact, a multitude

of DA strategies can be applied to dermoscopic images without altering their semantic content (Table 1). Although not all augmentation steps induce a performance boost, they can all be exploited during inference by feeding each test image multiple times to each network, and by subjecting each image to the same random DA process performed during training. By averaging the predictions of a single network, we employ an ensemble technique that increases both the overall accuracy and calibration without the need of training additional networks. The only drawback is an inference time overhead. This method can be especially beneficial when multiple networks are trained with diverse DA strategies and then merged together. In these cases, DA increases the ability of the framework to be robust to small transformation that are irrelevant towards the final diagnosis.

To take this one step further, we calibrate single models by means of Temperature Scaling [45], an easy to integrate calibration technique with good performance in image classification. Empirical results show that the Data Augmentation Ensemble (DAE) improves both the accuracy and the calibration of the final ensemble of CNNs, whereas the usage of Temperature Scaling, despite successfully lowering the Expected Calibration Error (ECE) of a single model, degrades the overall ensemble calibration while having a small impact on the

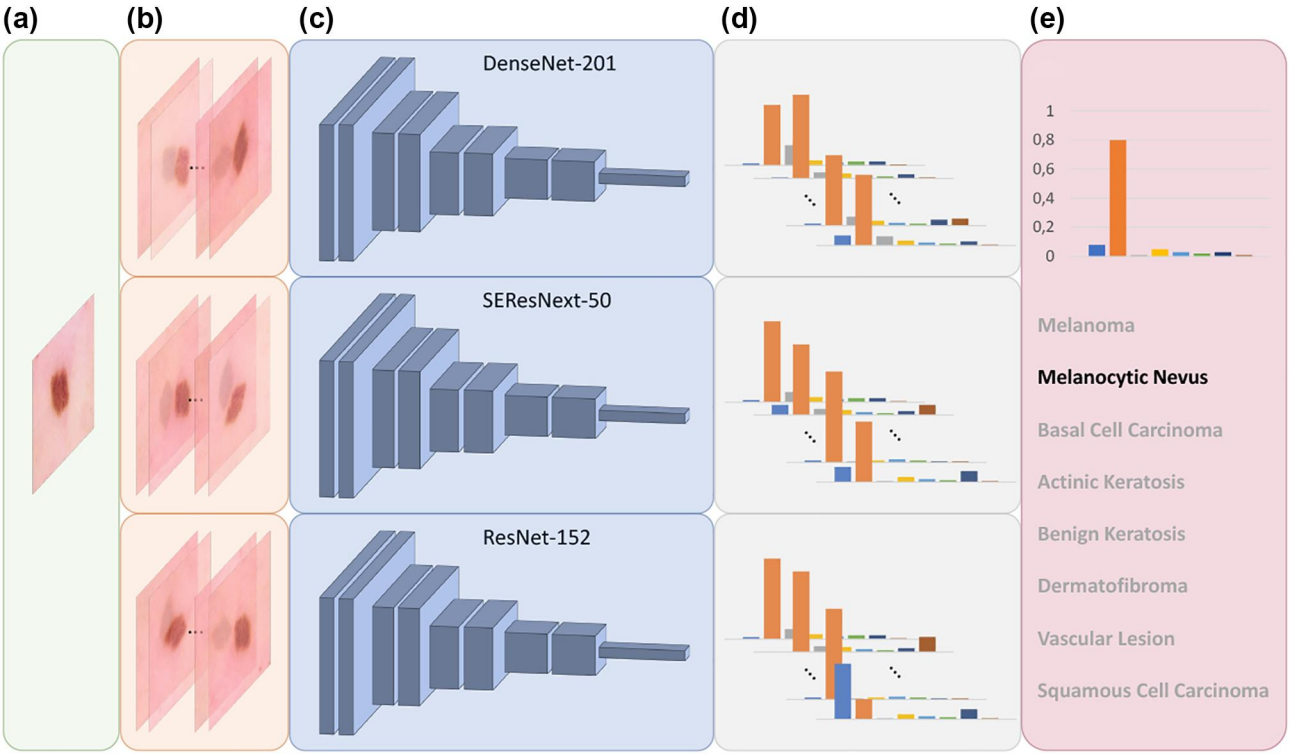

FIGURE 6 Illustration of the proposed ensemble method. Given an input image (a) at inference time, multiple versions of it are obtained by means of random data augmentation. Augmented images (b) are then fed to a convolutional neural networks (c) and calibrated outputs (d) are obtained. This process is repeated over multiple convolutional neural networks and, finally, all of the network outputs are averaged together in the last step in order to obtain the final prediction (e)Model calibration

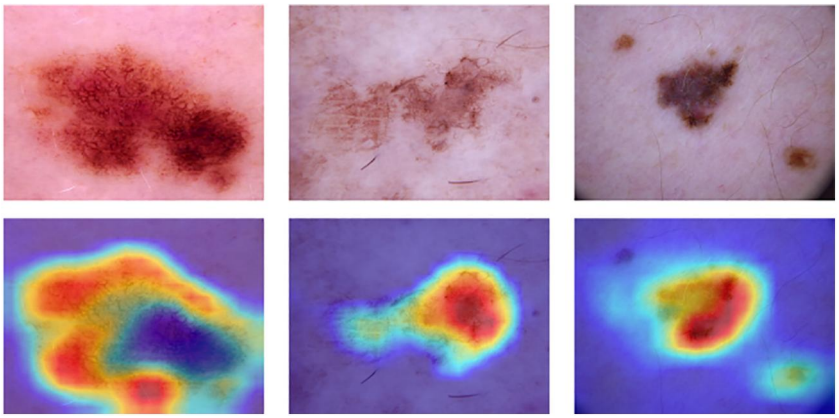

FIGURE 7 Samples from the 2019 international skin imaging collaboration dataset and attention heatmaps obtained by means of the Grad-CAM method. Besides always locating the skin lesion, the network can aim its attention at specific sections such as lesion borders (leftmost image) or darker patches (rightmost image)

TABLE 4 Results obtained through a calibrated ensemble using different DA configurations and fixed 512 × 512 image size over the training partition

Net	Data augmentation	Balanced accuracy	AUC
DenseNet-201	2	0.861	0.980
SEResNeXt-50	0	0.867	0.982
ResNet-152	3	0.863	0.979
Ensemble	–	0.888	0.988

Abbreviations: AUC, area under the roc curve; DA, data augmentation.

TABLE 5 Results obtained through an ensemble of the three top performing architectures described in [32]. The two input strategies employed in the article are RR cropping and SS cropping

Net	Input strategy	Balanced accuracy	AUC
EfficientNet-b4	RR	0.800	0.965
EfficientNet-b5	SS	0.777	0.957
EfficientNet-b6	SS	0.773	0.957
Ensemble	–	0.871	0.983

Abbreviations: RR,random-resize; SS, same-size.

TABLE 6 Confusion Matrix to show results of the calibrated ensemble detailed in Table 4, that is using different DA configurations and fixed 512 × 512 image size over the training partition. Each cell contains the percentage of images of a specific class defined by the column that were assigned to the class specified by the row. Accordingly, the main diagonal (green squares) displays the architecture sensitivity for each class

prediction \ ground truth	MEL	NV	BCC	AK	BKL	DF	VASC	SCC
MEL	85	4	1	3	3	0	1	1
NV	11	94	1	0	6	1	3	0
BCC	1	1	95	6	1	1	1	9
AK	1	0	1	84	3	1	0	7
BKL	2	1	1	6	85	0	0	3
DF	0	0	0	0	0	96	0	0
VASC	0	0	0	0	0	0	95	0
SCC	0	0	1	3	2	0	0	79

Abbreviation: DA, data augmentation.

accuracy. However, these results cannot yet be generalised for more sophisticated calibration techniques given in [40, 45, 49, 50]. We illustrate the whole proposed pipeline in Figure 6.

TABLE 7 BA and AUC obtained by ResNet architectures with different TPS and image resolutions

		512 × 512		256 × 256		128 × 128	
Network	TPS	BA	AUC	BA	AUC	BA	AUC
ResNet-18	10,000	0.699	0.952	0.690	0.945	0.625	0.918
ResNet-18	5000	0.610	0.907	0.617	0.913	0.560	0.893
ResNet-18	1000	0.431	0.856	0.450	0.863	0.416	0.826
ResNet-50	10,000	0.750	0.959	0.722	0.943	0.614	0.914
ResNet-50	5000	0.663	0.930	0.637	0.909	0.511	0.883
ResNet-50	1000	0.456	0.865	0.464	0.842	0.413	0.806
ResNet-152	10,000	0.773	0.964	0.739	0.951	0.648	0.918
ResNet-152	5000	0.689	0.936	0.633	0.925	0.572	0.889
ResNet-152	1000	0.522	0.891	0.495	0.863	0.433	0.831

Abbreviations: AUC, area under the ROC curve; BA, balanced accuracy, TPS, training partition sizes.

5 | EXPERIMENTAL RESULTS

This section presents the impact that different architectures, image resolutions, augmentation strategies, and dataset sizes have on the classification capabilities of several neural networks. Each network is trained using stochastic gradient descent (SGD) with momentum and a plateau learning rate scheduler. A validation set of 1000 images is used to monitor the accuracy of the model at each epoch and to apply the early stopping technique. Each network is tested on 5000 images and trained on a set of 19,331 images, using a cross entropy loss function which is weighted according to the inverse prior probability of each class. Ratios between classes are preserved in the training set as well as in both the validation and the test set. To avoid trivial comparisons, training times are computed on the same machine, equipped with one NVIDIA GTX 1080 Ti with 11 GB of memory.

To provide a qualitative visualisation of the experimental results, we make use of the Grad-CAM method [51] to show the attention heatmaps associated to one of our models. Figure 7 shows the specific part of the image on which the trained model focusses to make its prediction. On the other hand, quantitative results are expressed by means of two different metrics: Balanced Accuracy and AUC.

In Tables 2 and 3 and Figure 5, we analyse the performance of different networks when image resolution, network depth, and network width are scaled up. The results of this investigation reveal that image resolution is the most relevant hyperparameter for this task. As a matter of fact, the accuracy obtained by EfficientNet-b5 can be achieved by smaller networks like ResNet-50, by virtue of merely increasing the size of input images. Indeed, deeper versions of EfficientNet are unable to take advantage of their width, depth, and image resolution without any supplementary data, yielding worst accuracy results than those of EfficientNet-b4.

Tables 4 and 6 show the results obtained by merging three different networks, employing the ensemble technique

TABLE 8 BA and AUC obtained by ResNet-50 with different TPSs, DA configurations, and image resolutions. Indexes of DA are the same introduced in Table 1

		512 × 512		256 × 256		128 × 128	
TPS	**DA**	**BA**	**AUC**	**BA**	**AUC**	**BA**	**AUC**
19,000	1	0.804	0.973	0.754	0.958	0.626	0.923
19,000	2	0.837	0.976	0.781	0.963	0.704	0.939
19,000	3	0.846	0.977	0.770	0.960	0.660	0.921
10,000	1	0.725	0.956	0.664	0.931	0.557	0.900
10,000	2	0.730	0.959	0.705	0.938	0.589	0.905
10,000	3	0.743	0.964	0.708	0.935	0.617	0.909
5000	1	0.642	0.927	0.586	0.902	0.482	0.872
5000	2	0.667	0.926	0.635	0.911	0.516	0.872
5000	3	0.656	0.928	0.639	0.910	0.549	0.871
1000	1	0.432	0.845	0.413	0.837	0.337	0.767
1000	2	0.453	0.855	0.450	0.842	0.377	0.800
1000	3	0.459	0.869	0.457	0.847	0.406	0.800

Abbreviations: AUC, area under the ROC curve; BA, balanced accuracy, DA, data augmentation; TPS, training partition sizes.

described in the previous Sections and pictured in Figure 6. For the sake of providing a fair evaluation of the proposed method, we make use of the guidelines and the code provided by the winners of the 2019 ISIC challenge (DAISYLabs [32]), and build an ensemble using the three top performing architectures described in the article. We implement the preprocessing strategy detailed by the authors using the European Computer Vision Library (ECVL) [52]. Each model is trained and tested using the same data partition described at the beginning of this Section, and the results are displayed in Table 5. The first three rows of Tables 4 and 6 display the performance of neural networks when tested individually, through a single forward pass, and with no ensemble techniques applied at inference time. On the other hand, the last row of Table 6 (ensemble) presents the results obtained by merging the output of three CNNs after applying the two prediction strategies described in [32], both of which involve multiple forward passes for each network. The comparison between Tables 4 and 6 clearly shows that the proposed method outperforms the winners of the ISIC 2019 challenge, when the two algorithms are trained and tested using the exact same data.

To investigate how our framework performance is affected by the amount of available data, we randomly build three smaller version of the ISIC dataset with, respectively, 10,000,

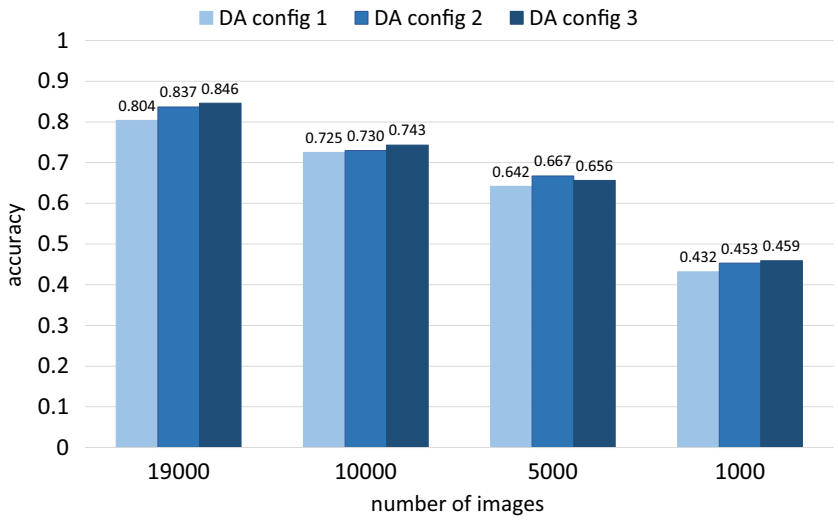

FIGURE 8 The effects that different dataset sizes have on ResNet-50. As more labelled data is available, the overall accuracy is always increased. For each dataset size, three different convolutional neural networks are trained using a different data augmentation strategy (Table 1), showing how adding data augmentation steps during training time usually boosts the accuracy

TABLE 9 Results obtained on the official 2019 ISIC Challenge [53], with a description of data used for every method. The last column indicates methods that employ OOD strategy

Method	**Balanced Accuracy**	**AUC**	**Employed Images**	**OOD**
DAISYLabs [32]	0.636	0.923	ISIC 2019 + additional data	✗
DysionAI	0.607	0.780	ISIC 2019	✓
Proposed Method	**0.593**	**0.886**	**ISIC 2019**	✗
DermaCode	0.578	0.892	ISIC 2019	✗
Nurithm Labs	0.569	0.870	ISIC 2019 + additional data	✗

Abbreviations: AUC,area under the ROC curve; ISIC, international skin imaging collaboration, OOD, out of distribution.

5000, and 1000 images, always preserving the ratio between different classes. Table 7 underlines that dataset size is a key factor for obtaining good performance, as reducing the amount of available data always worsens the final accuracy. Moreover, the impact of high resolution is only reduced for drastic configuration, with very shallow networks and extremely little data available. In Table 8, we also present the effects obtained by increasing the number of performed DA steps during the training process, with datasets of various sizes. Results of Tables 7 and 8 are summed up in Figure 8.

Finally, Table 9 displays the results obtained on the official 2019 ISIC challenge by following the guidelines defined in this article. The official metric of the challenge is the balanced accuracy, and AUC values are added for completeness. Metrics are computed over nine classes as described in Section 3, the *none of the others* class is not considered by our approach because no labelled data is provided.

The proposed ensemble strategy is the best performing method when compared with algorithms that do not employ an out of distribution detection technique to handle the ninth class and take advantage of no additional data.

6 | CONCLUSION

In this article, we addressed the impact of image resolution, DA, and different state-of-the-art architectures on dermoscopic images analysis. Furthermore, an ensemble strategy that considers augmented samples at test time is presented. Our method successfully deals with the absence of balance between classes, by means of a large use of DA strategies (both at training and testing time), and a weighted cross entropy loss. The proposed solution takes advantage of multiple networks trained using different augmentation methods. A probabilistic approach is employed to perform the ensemble over calibrated network decisions thus ensuring better results.

Carrying out experiments in a systematic way, we proved that dermoscopic image analysis highly depends on input image resolution and that the amount of performed DA strategies and available labelled data play a major role in this task. We empirically demonstrated that owing to the deficiency of dermoscopic labelled data with respect to natural images, extremely deep architectures (e.g. SEResNeXt-101, EfficientNet-b7) fail to provide better results than shallower ones, highlighting that conclusions on natural images cannot be directly extended to dermoscopic ones.

The ISIC dataset is expected to consistently keep growing in the approaching years, we therefore plan to extend our studies using larger amounts of data to assess how dermoscopic images analysis is altered if more data is available, if the class imbalance issue is stretched, or if the number of classes is increased. Moreover, state-of-the-art results are obtained by means of extremely expensive inference procedures (both in terms of time and hardware resources). Future research directions will hence include an investigation to find cheaper ways to obtain class predictions from neural networks without excessively lowering the achieved discrimination capabilities. Finally, we plan to explore the effects of new state-of-the-art calibration techniques on an ensemble of neural networks.

The proposed ensemble yielded a balanced accuracy of 0.593 on the official 2019 ISIC challenge, achieving the third best result. It is the best performing method when compared with the challenge participants that do not exploit additional data and do not take advantage of an Out of Distribution Data detector. To ensure reproducibility, the source-code is provided in [54].

ACKNOWLEDGEMENTS

Juan Maroñas is supported by grant FPI-UPV, grant agreement No 825,111 DeepHealth Project, and by the Spanish National Ministry of Education through grant RTI2018-098091-B-I00. The research leading to these results has received funding from the European Union through *Programa Operativo del Fondo Europeo de Desarrollo Regional* (FEDER) from *Comunitat Valencia* (2014-2020) under project *Sistemas de frabricación inteligentes para la indústria 4.0* (grant agreement IDI-FEDER/2018/025).

ORCID

Federico Pollastri https://orcid.org/0000-0001-8036-1559
Federico Bolelli https://orcid.org/0000-0002-5299-6351
Costantino Grana https://orcid.org/0000-0002-4792-2358

REFERENCES

1. Bray, F., et al.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J. Clin. 68(6), 394–424 (2018)
2. Rigel, D.S., Russak, J., Friedman, R.: The evolution of melanoma diagnosis: 25 years beyond the ABCDs. CA A Cancer J. Clin. 60(5), 301–316 (2010)
3. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29(6), 82–97 (2012)
4. Bolelli, F., Baraldi, L., Grana, C.: A hierarchical quasi-recurrent approach to video captioning. In: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), pp. 162–167 (2018)
5. He, K., et al.: Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2015)
6. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019)
7. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press Cambridge (2016)
8. Mercadante, C., et al.: A cone beam computed tomography annotation tool for automatic detection of the inferior alveolar nerve canal. In: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 1–8 (2021)
9. Codella, N.C., et al.: Skin lesion analysis towards melanoma detection: a challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 168–172 (2018)
10. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data. 5, 180161 (2018)

11. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
12. Bengio, Y., et al.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5(2), 157–166 (1994)
13. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift, pp. 1–11.arXiv preprint arXiv:150203167 (2015)
14. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of Machine Learning Research, pp. 249–256 (2010)
15. LeCun, Y.A., et al.: Efficient BackProp. In: Neural Networks: Tricks of the Trade, pp. 9–48 (2012)
16. Xie, S., et al.: Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1492–1500 (2017)
17. Huang, G., et al.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
19. Deng, J., et al.: A large-scale hierarchical image database'. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
20. He, K., Girshick, R., Dollár, P.: Rethinking ImageNet pre-training. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4918–4927 (2019)
21. Raghu, M., et al.: Understanding transfer learning for medical imaging'. In: Advances in Neural Information Processing Systems, pp. 3342–3352 (2019)
22. Harangi, B.: Skin lesion classification with ensembles of deep convolutional neural networks. J. Biomed. Inf. 86, 25–32 (2018)
23. Al.Masni, M.A., et al.: Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. Comput. Methods Progr. Biomed. 190, 105351 (2020). https://doi.org/10.1016/j.cmpb.2020.105351
24. Valle, E., et al.: Data, depth, and design: learning reliable models for skin lesion analysis. Neurocomput. 383, 303–313 (2020)
25. Ligabue, G., et al.: Evaluation of the classification accuracy of the kidney biopsy direct immunofluorescence through convolutional neural networks. CJASN. 15(10), 1445–1454 (2020)
26. Allegretti, S., et al.: Supporting skin lesion diagnosis with content-based image retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 1–8. (2021)
27. Hu, Z., et al.: Deep learning for image-based cancer detection and diagnosis − a survey. Pattern Recogn. 83, 134–149 (2018)
28. Yuan, Y., Chao, M., Lo, Y.-C.: Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance. IEEE Trans. Med. Imag. 36(9), 1876–1886 (2017)
29. Barata, C., Celebi, M.E., Marques, J.S.: Explainable skin lesion diagnosis using taxonomies. Pattern Recogn. 110, 107413 (2020)
30. Wang, X., Ding, H., Jiang, X.: Dermoscopic image segmentation through the enhanced high-level parsing and class weighted loss. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 245–249. (2019)
31. Wang, X., et al.: Bi-directional dermoscopic feature learning and multi-scale consistent decision fusion for skin lesion segmentation. IEEE Trans. Image Process. 29, 3039–3051 (2020)
32. Gessert, N., et al.: Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX. 7, 100864 (2020). https://doi.org/10.1016/j.mex.2020.100864
33. Combalia, M., et al.: BCN20000: dermoscopic lesions in the wild, pp. 1–3.arXiv preprint arXiv:190802288 (2019)
34. Whited, J.D., Grichnik, J.M.: Does this patient have a mole or a melanoma? JAMA. 279(9), 696–701 (1998)
35. Perez, F., et al.: Data augmentation for skin lesion analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 303–311 (2018)
36. Pollastri, F., et al.: Improving skin lesion segmentation with generative adversarial networks. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pp. 442–443.(2018)
37. Pollastri, F., et al.: Augmenting data with GANs to segment melanoma skin lesions. Multimed. Tool Appl. 79(21–22), 15575–15592 (2019)
38. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout, pp. 1–8.arXiv preprint arXiv: 170804552 (2017)
39. MacKay, D.J.C.: The evidence framework applied to classification networks. Neural Comput. 4(5), 720–736 (1992)
40. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, pp. 6402–6413 (2017)
41. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer-Verlag (2006)
42. Kim, H.C., Ghahramani, Z.: Bayesian classifier combination. In: Lawrence, N.D., Girolami, M. (eds.) Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. Vol. 22 of Proceedings of Machine Learning Research, pp. 619–627 (2012)
43. Gessert, N., et al.: Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 67(2), 495–503 (2019)
44. Pollastri, F., et al.: Confidence calibration for deep renal biopsy immunofluorescence image classification. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 1–8 (2021)
45. Guo, C., et al.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Vol. 70 of Proceedings of Machine Learning Research, pp. 1321–1330 (2017)
46. DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. The Statistician. 32, 12–22 (1983)
47. Canalini, L., et al.: Skin lesion segmentation ensemble with diverse training strategies. In: International Conference on Computer Analysis of Images and Patterns, pp. 89–101 (2019)
48. Izmailov, P., et al.: Averaging weights leads to wider optima and better generalization, pp. 1–12. arXiv preprint arXiv:180305407 (2018)
49. Kumar, A., et al.: Trainable calibration measures for neural networks from kernel mean embeddings. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Vol. 80 of Proceedings of Machine Learning Research, pp. 2805–2814 (2018)
50. Maroñas, J., Paredes, R., Ramos, D.: Calibration of deep probabilistic models with decoupled Bayesian neural networks. Neurocomput. 407, 194–205 (2020)
51. Selvaraju, R.R., et al.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
52. Cancilla, M., et al.: The DeepHealth Toolkit: a unified framework to boost biomedical applications. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 1–8 (2021)
53. ISIC 2019 Leaderboard, ISIC-Archive, 2019 Accessed 5 Jan 2021. https://challenge.isic-archive.com/leaderboards/2019
54. Pollastri, F., et al.: Enleadertwodots 'Paper Source Code', GitHub (2020). Available from https://github.com/PollastriFederico/ISIC2019. Accessed 5 Jan 2021.

How to cite this article: Pollastri, F., et al.: A deep analysis on high-resolution dermoscopic image classification. IET Comput. Vis. 15(7), 514–526 (2021). https://doi.org/10.1049/cvi2.12048