

## RESEARCH ARTICLE

WILEY

# Handling dataset dependence with model ensembles for skin lesion classification from dermoscopic and clinical images

Ellák Somfai<sup>1,2</sup>  | Benjámín Baffy<sup>1</sup>  | Kristian Fenech<sup>1</sup>  | Rita Hosszú<sup>3</sup> |  
Dorina Korózs<sup>3</sup>  | Marcell Pólik<sup>1</sup>  | Miklós Sárdy<sup>3</sup>  | András Lőrincz<sup>1</sup> 

<sup>1</sup>Department of Artificial Intelligence, Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary

<sup>2</sup>Institute for Solid State Physics and Optics, Wigner Research Centre for Physics, Budapest, Hungary

<sup>3</sup>Department of Dermatology, Venereology and Dermatocology, Faculty of Medicine, Semmelweis University, Budapest, Hungary

## Correspondence

Ellák Somfai, Department of Artificial Intelligence, Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary.  
Email: [somfaiellak@inf.elte.hu](mailto:somfaiellak@inf.elte.hu)

## Present address

Rita Hosszú, National Institute of Mental Health, Neurology and Neurosurgery, Nyíró Gyula Hospital, Budapest, Hungary.

## Funding information

European Social Fund; Nemzeti Kutatási Fejlesztési és Innovációs Hivatal; National Research, Development and Innovation Fund of Hungary, Grant/Award Number: TKP2020-NKA-06; Hungarian Ministry of Innovation and Technology

## Abstract

Dataset dependence affects many real-life applications of machine learning: the performance of a model trained on a dataset is significantly worse on samples from another dataset than on new, unseen samples from the original one. This issue is particularly acute for small and somewhat specific databases in medical applications; the automated recognition of melanoma from skin lesion images is a prime example. We document dataset dependence in dermoscopic skin lesion image classification using three publicly available medium size datasets. Standard machine learning techniques aimed at improving the predictive power of a model might enhance performance slightly, but the gain is small, the dataset dependence is not reduced, and the best combination depends on model details. We demonstrate that simple differences in image statistics account for only 5% of the dataset dependence. We suggest a solution with two essential ingredients: using an ensemble of heterogeneous models, and training on a heterogeneous dataset. Our ensemble consists of 29 convolutional networks, some of which are trained on features considered important by dermatologists; the networks' output is fused by a trained committee machine. The combined International Skin Imaging Collaboration dataset is suitable for training, as it is multi-source, produced by a collaboration of a number of clinics over the world. Building on the strengths of the ensemble, it is applied to a related problem as well: recognizing melanoma based on clinical (non-dermoscopic) images. This is a harder problem as both the image quality is lower than those of the dermoscopic ones and the available public datasets are smaller and scarcer. We explored various training strategies and showed that 79% balanced accuracy can be achieved for binary classification averaged over three clinical datasets.

## KEYWORDS

deep learning, skin lesion classification

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *International Journal of Imaging Systems and Technology* published by Wiley Periodicals LLC.

## 1 | INTRODUCTION

Dataset dependence is a difficulty often encountered in real-life machine-learning applications, and it is particularly bothersome in the medical domain. The phenomenon manifests itself in the observation that when a machine learning model is trained on dataset *A*, its performance on the hold-out validation part of the same dataset (or better still, its evaluation on dataset *A* via cross-validation) is better than its performance on dataset *B*; and vice versa, when a model is trained on dataset *B*, its (cross-validation) performance on dataset *B* is better than on dataset *A*. The symmetrical performance drop implies that neither database is “more difficult” than the other. Instead, we assume that distributions of the samples of the two datasets do not completely overlap in some abstract feature space, even though they should sample the same overall distribution. The high-performance deep neural networks may have poor results during testing if the training set does not properly represent the test set.

Dataset dependence is known in diverse domains, including, for example, lung cancer, where considerable efforts have been made to develop reproducible machine learning methods.<sup>45</sup> Similar problems are unavoidable in skin lesion classification, but controlled studies are lacking.

There are many reasons why datasets can differ. A simple source is differences in data acquisition: lighting condition, the quality and uniformity (e.g., bubbles) of the optical contact fluid, the relative proportion of the lesion compared to the saved image frame, image sharpness, and presence of device artifacts like millimeter scale marks, among others. However, there are more subtle issues, for example, ethnic differences not only affect skin color (which in itself affects model performance) but also give rise to differences in the distribution of the affected anatomic site<sup>10</sup>; this difference is also observed in alternative pathways to melanoma.<sup>14</sup> Melanoma, the uncontrolled proliferation of melanocytes manifests itself in diverse forms (see e.g.,<sup>39</sup>). Its features may also change due to population genetic and geographic differences<sup>44</sup> constraining sample collections.

To explore dataset dependence and methods to reduce it on a concrete example, we consider the binary classification of dermoscopic skin lesion images as melanoma versus anything else. The recognition of melanoma is highly motivated due to medical and public health reasons: while the worldwide incidence is increasing for both melanoma<sup>43</sup> and non-melanoma skin cancer types,<sup>28</sup> malignant melanoma is the most aggressive of them, causing the highest absolute number of deaths of skin cancers. The 5-year survival rate of melanoma is

99% if caught early, underlining the importance of early detection. Automated recognition of melanoma has a strong potential to contribute to this effort, especially due to the recent upsurge of demand for teleradiology, caused, for example, by the COVID-19 pandemic among other reasons.

In recent years, however, mainly due to the widespread and low-cost availability of high-resolution digital imaging, for example, with mobile phones, demand has increased to identify malignant skin lesions from plain, non-dermoscopic photos. These images, often called *clinical* or *macro* photos, are potentially well-suited for remote and automated diagnosis. Such applications are challenged by two difficulties: (i) the quality of macro images is lower, as the light reflection from the skin surface makes it harder to identify features under the skin surface, and (ii) the size and availability of high-quality public macro image sets is limited compared to the dermoscopic counterpart. On the other hand, automated analysis of macroscopic images could be of great value for many potentially concerned people in remote areas, for example.

In this paper, we focus on the ultimate goal of classifying melanoma versus non-melanoma skin lesion images coming from *an unknown source*. Because of dataset dependence, this task is harder than designing a model which performs well on a hold-out test set of a training dataset—which is the objective of the majority of skin lesion classification papers in the literature. Our choice of training and testing sets, as well as the selection and design of models (ensembles), are motivated by this ultimate goal.

Our contributions in this paper are as follows:

- We demonstrated dataset dependence for skin lesion classification using three roughly equal size publicly available mono-source (each dataset is from a single clinic) dermoscopic image datasets.
- We explored whether common techniques used with deep neural networks improve the generalization power of a skin lesion classifier.
- We showed that generic image-level differences, including the distribution of the relative size of the lesion within the image account for a small fraction (we observed only 5%) of the dataset dependence, the rest must be due to more subtle dissimilarity.
- We showed that a successful strategy to tackle dataset dependence is twofold: use an ensemble with heterogeneous models, and train it with a multi-source dataset. We built a novel model ensemble consisting of 29 convolution networks from three model families, which differ even in the training target and the input image. A number of the constituent networks are especially

suitable for taking part in a melanoma classification ensemble, as they are either trained to recognize a physical feature relevant to melanomas or focus on parts of the lesion image that are considered important to recognize melanoma by dermatologists. As the training target is not identical across the constituent models, conventional aggregation methods like averaging or majority voting are not suitable; we used a shallow neural network instead. For training, the multi-source combined International Skin Imaging Collaboration (ISIC) dataset was used.

- We applied the model ensemble for a harder problem of classifying clinical skin lesion images. Determined the best training strategy, and achieved 79% balanced accuracy for binary classification averaged on three datasets.

## 2 | RELATED WORKS

Significant progress has happened in the last 6–7 years in automated skin lesion classification using machine learning techniques (see<sup>30</sup> for a quality-filtered, systematic review). The development was fueled by the increased availability of high-quality skin lesion datasets (see Section 3.1), and the development of general-purpose image classification tools, especially deep convolutional networks (convnets) based methods,<sup>4,19,46</sup> sparse coding,<sup>6</sup> and more recently attention based learning.<sup>47</sup> It is remarkable, that a deep residual network (in particular, a classifier based on the ResNet50 backbone) is capable to outperform dermatologists in classifying melanoma and atypical nevi from dermoscopic images under suitable conditions.<sup>3</sup>

The challenges organized by the ISIC every year since 2016 contributed to focusing the efforts by practitioners to develop better and better methods. Recently, all top performers of the classification challenge of 2019 applied deep learning methods, and in some cases, ensembles of deep neural networks were used. Gessert et al.<sup>13</sup> trained multiple variants of different convnets, predominately using EfficientNet architectures. These networks are trained at different resolutions and the final prediction is obtained by a weighted sum over the classifiers. In this case, no special attention was given to particular features of the images apart from cropping the lesion.

In the most recent challenge, in 2020 the winning method used an ensemble of 18 convnets, where the model variability was achieved by using a variety of backbone networks.<sup>16</sup> They adjusted the input image size such that the performances of the constituent networks were comparable, so combining them by simple averaging worked well. The convnets employed are very strong,

which was also reflected in the very substantial computational resources needed to train them.

The most recent developments since the year 2021 include the combination of deep convnets with bidirectional short- and long-term memory networks<sup>1</sup>; the combination of mask-based region of interest cropping, classification based on convnets, and class balancing<sup>17</sup>; using distributed densely connected convnets<sup>41</sup>; hierarchical arrangement of convnet-based classifiers<sup>2</sup>; comparison of different architectures<sup>11</sup>; and careful training of a light architecture, usable in a mobile environment as well.<sup>21</sup>

Prior to the surge in deep learning-based methods, work focusing on the classification of melanoma and nevi using conventional statistical learning methods had been explored. Melanoma and nevi were distinguished by local feature extraction,<sup>32</sup> which takes into account such attributes as lesion shape, border asymmetry, color, and texture variations. The authors used the extracted features with different classifiers. In more recent work, seven hand-selected features of skin lesions indicative of melanoma were used,<sup>20</sup> such as the presence of a blue-whitish veil, or the type of pigment network and vascular structures if they are present. Separate classifiers were trained for these features on a dataset containing detailed ground truth values for the features. These were combined by a weighted sum of the post-thresholded predictions: the binary predictions of the feature classifiers were added with a weight of either 2 or 1, and the sample was considered positive if the score reached either 1 or 3 (based on whether the focus was on sensitivity or specificity). Since a number of the features have strong predictive power (e.g., the blue-whitish veil in itself predicts melanoma with a specificity of 97%, although with only 51% sensitivity<sup>26</sup>), we incorporated them in our ensemble (Section 3.3) using an improved data fusion mechanism.

Previous work on clinical images is typically limited to small datasets, where an accuracy of 0.8 for binary melanoma detection was achieved.<sup>31</sup> Larger investigations tend to classify multiple skin lesion types,<sup>23</sup> where the identification of melanoma is less prominent.

In the present work, instead of simple or weighted averaging we employ a trained method—a shallow net committee machine—which is capable of handling models with differing performances and complementary strengths and fusing the above-mentioned achievements. Model components contain potent general-purpose image classifiers trained to distinguish melanoma from other skin lesions, as well as classifiers focusing on specific features. This way we exploit human efforts that went into identifying the decisive attributes, hoping that (a) focusing on the presence or absence of the relevant features may increase robustness and (b) these features

may overcome—at least partially—the limited amount of data in some datasets.

### 3 | MATERIALS AND METHODS

#### 3.1 | Datasets

In the experiments, we use a number of publicly available datasets, and a private one.

HAM10000 is a well-documented dataset<sup>40</sup> containing little over 10 000 dermoscopic images of skin lesions, intended for benchmarking machine learning models. The dataset is curated, high quality, but multi-source: it contains four sub-datasets from two clinics. For the purpose of systematic evaluation of melanoma versus non-melanoma classification, we chose two sub-datasets: Vidir\_modern (VM) and Rosendahl (ROS), as they are similar in size and melanoma fraction (15%–20%). For both datasets, we kept only a single image per lesion, ending up with 1695 and 1552 images respectively.

The publicly available ISIC challenge datasets from years 2018 (which includes HAM10000),<sup>7,40</sup> 2019,<sup>8,9,40</sup> and 2020<sup>33</sup> are large, high-quality multi-source datasets intended for multiple tasks. We use the combination of their diagnosis classification datasets (C-ISIC), where the duplicates are attempted to be included only once.<sup>16</sup> The segmentation dataset of ISIC2018 was used to train the lesion segmenter.

PH2 is a publicly available dataset obtained at the Hospital Pedro Hispano, Matosinhos, Portugal.<sup>25</sup> It contains dermoscopic images of 200 melanocytic lesions in total: 80 common nevi, 80 atypical nevi, and 40 melanomas, together with manual segmentation of the skin lesion, and a number of dermoscopic features.

The companion dataset of the Derm7pt method<sup>20</sup> is available publicly; it contains image pairs of 1011 lesions. All the lesions have both a dermoscopic image and a clinical (normal or macro) photograph. About 25% of the lesions are melanoma. In addition to histology-based classification and patient metadata, the dataset contains annotation for the presence of the seven visual features (per lesion) forming the derm7pt scheme. We use both the dermoscopic images (DERM7D) and the clinical (macro) images (DERM7M).

PAD-UFES: a public dataset from the Federal University of Espírito Santo (UFES) in Brazil. The dataset contains 2298 clinical (macro) images recorded with smartphones categorized into six disease types, accompanied by metadata of the lesion's features and various demographics about the patients. The dataset only accommodates 50 melanoma images, making it very unbalanced.

MED-NODE: a publicly available dataset originating from the Dermatology department of the University of Groningen<sup>15</sup> containing 100 nevus and 70 melanoma clinical images.

Semmelweis University dataset: a private set of images, collected by the Department of Dermatology, Venereology and Dermatooncology of Semmelweis University, Budapest. From the collection of the University, we tested a selection of 156 images (curated to maintain consistent image quality and variability), of which 33% were melanoma (Semmelweis).

#### 3.2 | Single models

To evaluate dataset dependence, and especially to study how it varies with the details of the model, we employed classifiers developed for the separation of melanoma and non-melanoma images. For baseline, we used a classifier with a medium strength backbone: EfficientNet level b4.<sup>36</sup> For improved representation, the model was trained for 9-way classification for the skin lesion categories of the ISIC2019 dataset, and the output was subsequently mapped to 2 classes (melanoma vs. anything other, which included other cancerous types).

Performance and dataset dependence were evaluated for different versions of this model when we explored techniques intended to improve the generalization power of deep convnets. The variations included (a) class-weighted training: the amplitude of the convnet's weight adjustments during backpropagation was inversely proportional to the class abundance, to counteract the effects of imbalanced training sets; (b) weight regularization, either implemented as traditional weight decay, or was decoupled from the gradient-based update<sup>24</sup> using the Adam optimizer (often called AdamW), and (c) heavy augmentation at training. To assess backbone dependence a more recent convnet was also tested: EfficientNetv2-M from 2021.<sup>37</sup> The setup of the classifiers was inspired by some of the constituent models of the award-winning algorithm developed in the SIIM-ISIC Melanoma Classification Challenge.<sup>16</sup> Technical details about the models are in the Appendix.

#### 3.3 | Ensemble models

To improve performance we employed ensembles of models: we used different sets of constituent models and derived results by combining their outputs in different ways.

The constituent models included the original 18 models of the winning solution of the 2020 SIIM-ISIC

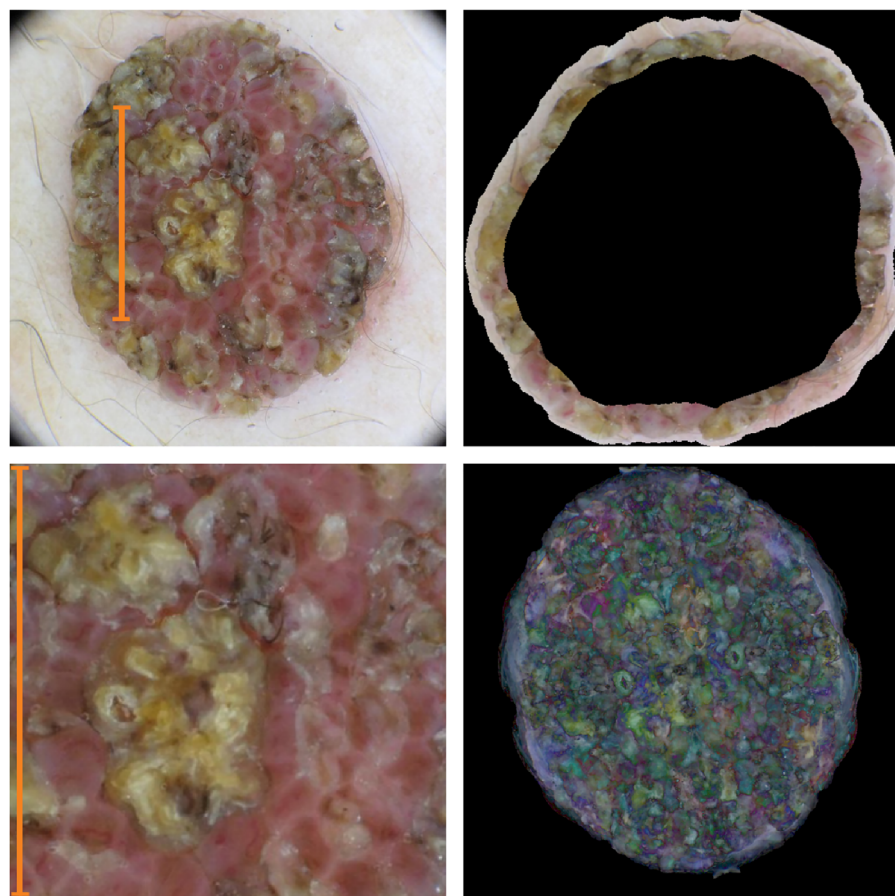


skin lesion classification challenge.<sup>16</sup> These are typically based on the EfficientNet family backbones ranging from b3 to b7 plus a ResNeSt model and a Squeeze-and-Excitation-ResNeXt (SE-ResNeXt) model. The classifiers are trained similarly to the single model of Section 3.2, with heavy augmentation but without class-weighted training. To increase model variability the input image size for the different models ranged between  $384 \times 384$  pixels and  $896 \times 896$  pixels, and four of the models used patient metadata as well, including sex and age. The training of these nets is hugely resource-intensive; therefore, the weights trained by<sup>16</sup> have been employed.

The next three convnets, which we call *feature classifiers*, were purpose-developed by us<sup>34</sup> to focus on features considered important to identify melanoma by dermatologists, in particular the widely used ABCDE rule and Mensies' list.<sup>27</sup> The specific information is selected by a suitable preprocessing of the input image (see Figure 1 for illustration), which is then fed to a classifier described in Section 3.2 with class-weighted training and heavy augmentation. The models are the *border* classifier (the closest match to the "B" Border criterion of the ABCDE rule), keeping only a narrow strip of the lesion's image around its perimeter; the *color asymmetry* classifier, designed to capture the asymmetry (criterion "A") in color and structure of the lesion, by using the RGB color

differences between the original image and its reflections; and the *central* classifier, intending to capture fine details of the lesion, such as brown or gray dots, globules and atypical pigment networks, by using a region around the centroid of the lesion at full resolution. The above models depend on the image mask of the lesion, which is computed by a U-net-based convnet, where the skip connections have been replaced by a channel attention block called "efficient channel attention"<sup>42</sup> to improve performance. More details about the feature classifiers and the segmenter can be found in.<sup>34</sup>

The last seven models, the *derm7pt classifiers* are also feature-based classifiers that target the criteria of the 7-point checklist.<sup>20</sup> These models, unlike the previous ones, are not trained on lesion diagnostic categories, but instead on detecting features potentially indicative of melanoma. The features are a pigment network, blue-whitish veil, vascular structures, pigmentation, streaks, dots and globules, and regression structures. Since—to our best knowledge—only the dataset accompanying the work of<sup>20</sup> contains relevant ground truth labels, the derm7pt classifiers can be trained on this single dataset only. Each feature is classified into 2–8 categories ("labels"), of which some are indicative of melanoma. For example, "pigment network" can be absent, typical, and atypical, and only atypical suggests melanoma. The



**FIGURE 1** Preprocessed input images for the feature classifiers. The images from left to right, then top to bottom: original image; border-image (rotated, see text); center image; and input to color asymmetry classifier. The contrast for the color asymmetry image has been increased on this figure for better visibility. The thin vertical orange scale bars on the original and center image correspond to the same physical length scale to illustrate the relation between the images.

classifiers are trained on these categories, which are then mapped to the melanoma-indicative binary classes.

The ensembles used in this work employ either 18 models (the constituent models of<sup>16</sup>), 22 models (the previous 18 models, plus our baseline model with class-weighted training and heavy augmentation, and the three feature classifiers), or 29 models (the previous 22 models plus the seven derm7pt classifiers).

### 3.4 | Committee machine

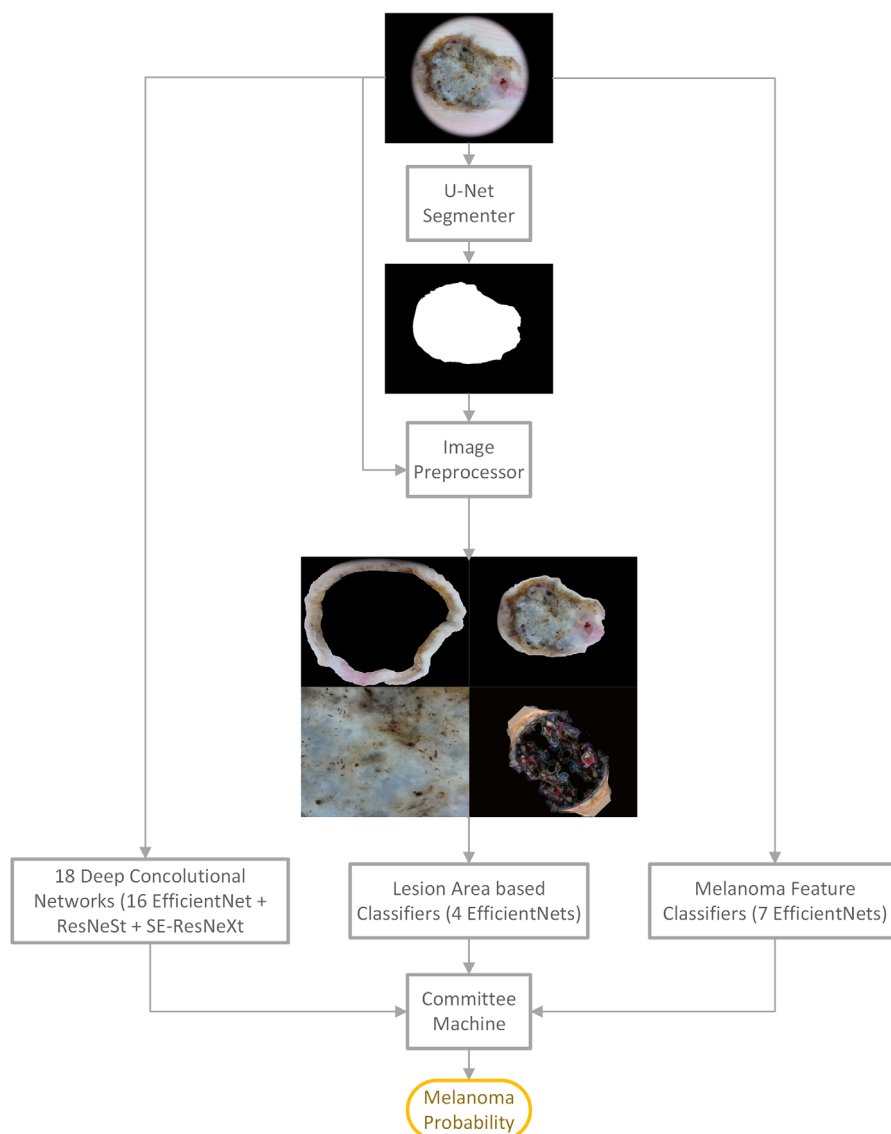
The simplest way to combine the models is by taking the *average* of their softmax outputs. This method was used on the 18-model ensemble<sup>16</sup> to win the 2020 SIIM-ISIC skin lesion classification challenge, therefore we consider this combination as the 2020 state of the art. A slightly more sophisticated way is to *average the calibrated probabilities* of each model, see Section 3.6 for more details.

Finally, the most flexible way to combine is a trained method, which we call a *committee machine*,<sup>38</sup> based on a shallow neural net. A visual overview of the 29-model ensemble combined with the committee machine is depicted in Figure 2.

We tried a number of different architectures for the committee machine and found the following optimal: three fully connected hidden layers consisting of 128, 64, and 32 neurons, with a dropout of 0.5 after each layer. During training, the categorical cross-entropy loss was optimized by the Adams optimizer with a learning rate of  $10^{-4}$  over 30 epochs using mini-batch size 192.

### 3.5 | Sharpness aware minimization

Recently a new regularization technique has been introduced for training artificial neural networks: instead of simply minimizing the loss, it is beneficial to minimize



**FIGURE 2** Overview of the architecture of our largest ensemble method (Committee-29)

both the loss value and the loss sharpness<sup>12</sup>: the method prefers network parameters that have *uniformly* low loss in the parametric neighborhood to improve the generalization capabilities of the network. This method, called “Sharpness Aware Minimization” (SAM) has the potential of improved prediction power, at the cost of longer training time. We experimented using this method, and its adaptive version (ASAM),<sup>22</sup> to train both the deep convnets and the committee machine.

### 3.6 | Calibrated probability

The raw (softmax) output of a neural network classifier trained on cross-entropy loss reflects the decision and the network’s confidence, but the actual value cannot be interpreted directly as probability, for example when training naively on unbalanced datasets. It is natural to expect that plotting the ground truth positive fraction of samples against the model’s raw prediction could provide a tool to obtain calibrated probabilities. Such plots have been known in the probabilistic forecasting literature for decades as reliability diagrams.<sup>5,18,29</sup>

We compute the calibrated probability, which for a given raw network output estimates the fraction of the samples that are ground truth positive (and elicit this network output). Details are provided in the Appendix.

## 4 | RESULTS

### 4.1 | Empirical comparison of dataset images

We display experiments using three independent datasets: ROS, VM, and DERM7D. These datasets are of comparable size (between 1000 and 1700 lesions per dataset), and in principle are expected to sample the same distribution: dermoscopic images of skin lesions obtained in clinical settings, containing 15%–25% melanoma. As we will see, despite the assumed similarity dataset dependence is observed; so it is instructive to check if a simple empirical comparison shows any difference.

In Figure 3, we show the histogram of the lesions compared to the full image area. Dataset VM contains more small lesions compared to ROS and DERM7D, the latter two have a similar distribution.

In Figure 4, the color composition of the dataset images is displayed. For a more appropriate color representation, the images were converted to the HSV color space. The histograms show that datasets VM and ROS have similar color distribution, while DERM7D is slightly different: the dominant hue is shifted towards orange

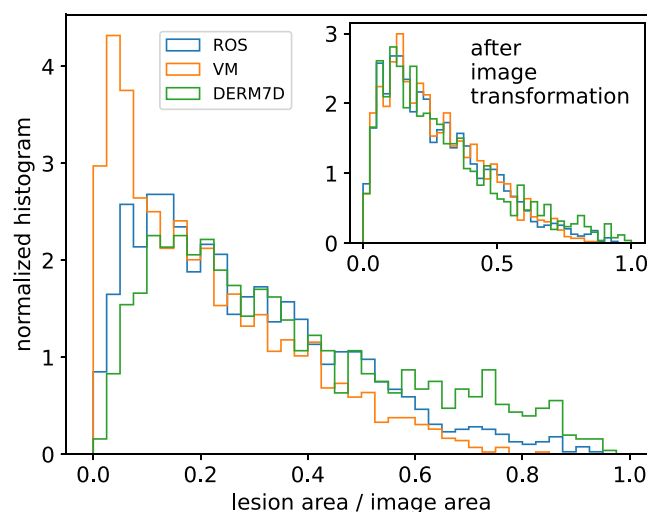
from red, the saturation has a wider distribution, and the images are brighter.

As we have seen, the three datasets differ in simple statistical properties. Moreover, in one of the datasets (DERM7D), a significant number of images contain black stripes of varying width at the image edges. To exclude the possibility that dataset dependence is caused by these simple differences, we transformed the VM and DERM7D image sets to match the statistical properties of ROS: (1) equalized the fraction histograms of the lesion areas, (2) removed black edges from DERM7D, and (3) equalized the color component histograms; see the Appendix for details. The resulting histograms are shown on the insets of Figures 3 and 4, respectively. Model performances on the transformed images will be shown in the next Section 4.2. The detailed discussion and interpretation of the results are compiled in Section 5.

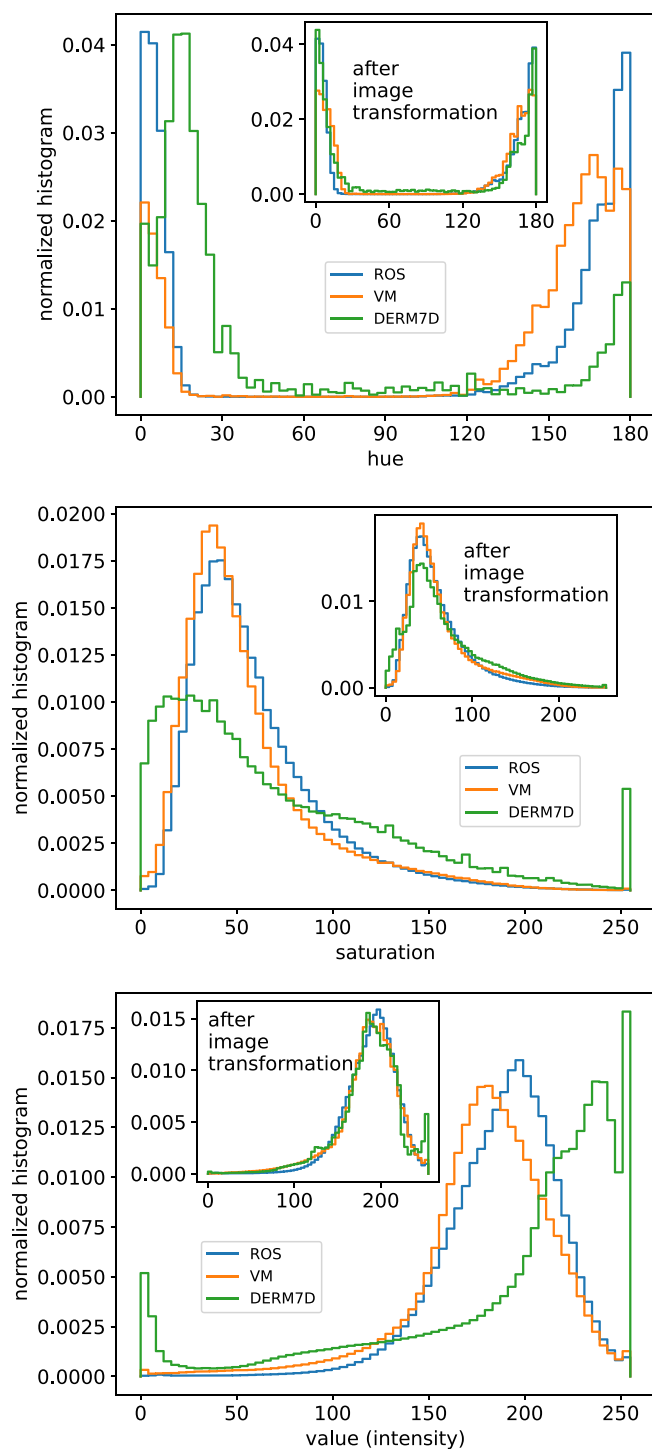
### 4.2 | Dataset dependence experiments on mono-source datasets

In the rest of this paper in all our experiments we considered the binary classification of the images as melanoma versus anything else, and employed five-fold cross-validation to decrease the evaluation noise. The evaluation of the training datasets was achieved via out-of-fold predictions, while on other, non-training datasets we averaged the output of the five models obtained from each fold.

To investigate dataset dependence, the following experiments were performed. For a given version of the



**FIGURE 3** Histogram of the fraction of the skin lesion area to the full image area. Inset: the same histogram, after image transformation of the VM and DERM7D datasets, see text for details.



**FIGURE 4** Histograms of the color components: hue, saturation, and value (intensity). The insets show the same histograms after the image transformation of the VM and DERM7D image sets, see text.

model, we trained on one of the three roughly equal size medium size mono-source datasets (ROS, VM, and DERM7D), and tested on all three datasets separately (altogether 9 measurements per model version).

Balanced accuracy calculated (a) from the raw softmax values and (b) from calibrated probabilities are shown in Table 1 for a range of model versions, evaluated both for the training set and for the other datasets. For the EfficientNet-b4 (EffNet-b4 for short) backbone, some of the techniques, in particular, the use of both class-weighted training and heavy augmentation resulted in a slight improvement. These cases are displayed as the first three bar groups in Figure 5. The figure demonstrates that in all cases models perform better on the training dataset than on other datasets and that dataset dependence is strong. The best combination for EfficientNet-b4 and EfficientNet-v2-M (EffNetv2-M for short) was achieved by different combinations of the applied techniques.

We studied whether the dataset dependence is caused by differences in simple image statistics. We checked the performance of the best model, that is, on EffNet-b4 + CWT + AUG, where CWT and AUG denote class-weighted training and heavy augmentation, respectively. We modified the datasets: VM and DERM7D images were transformed to match the statistical properties of ROS, see Section 4.1. As expected, performance on the training sets did not change, but on the other datasets a slight improvement was seen, especially when equalizing the lesion area fraction distribution, and fixing the edges. However, the improvement is small, amounting to only about 5% of the dataset dependence gap (Table 2).

### 4.3 | Training on a large multi-source dataset

We performed further experiments on the best (EffNet-b4 + CWT + AUG) version of the model: training was performed on the combined ISIC dataset, which is much larger. However, datasets VM and ROS are subsets of these large ISIC datasets as they are included in HAM10000. In turn, we took advantage of (a) the medium-sized DERM7D dataset and (b) two smaller datasets, namely PH2 and the private Semmelweis dataset. The employment of a large training set resulted in a large improvement when both evaluated on the training dataset, and also on the other datasets; see Table 3 for quantitative numbers. The dataset dependence, in this case, is remarkably small for averages of softmax-based balanced accuracy (filled circles), as clearly visible in the last bar group of Figure 5 as well. Database dependence drops to about one-fifth of the changes shown in the first three bar groups. However, averages of calibrated probability-based balanced accuracy decrease only slightly, by about 20%, or so.



**TABLE 1** ROS, VM, and DERM7D datasets: Balanced accuracy, and balanced accuracy evaluated on the calibrated probabilities (BA on Cal.Pr.), both at decision threshold 0.5

Model	Balanced accuracy		Balanced Acc. on Cal. Pr.	
	Train	Other	Train	Other
<b>EffNet-b4 (baseline)</b>	73.7 ± 0.46	59.8 ± 0.34	79.6 ± 0.27	67.6 ± 0.51
<b>EffNet-b4 + CWT</b>	74.9 ± 0.38	60.4 ± 0.28	79.9 ± 0.44	67.0 ± 0.75
EffNet-b4 + CWT + WD	74.7 ± 0.49	60.5 ± 0.48	80.5 ± 0.41	67.6 ± 0.51
<b>EffNet-b4 + CWT + AUG</b>	77.0 ± 0.39	<b>61.8 ± 0.22</b>	80.3 ± 0.68	<b>67.9 ± 0.56</b>
EffNet-b4 + CWT + AUG + WD	77.0 ± 0.42	61.4 ± 0.43	80.2 ± 0.42	67.6 ± 0.31
EffNet-b4 + CWT + AUG + WD*	77.0 ± 0.26	61.3 ± 0.37	80.0 ± 0.82	67.7 ± 0.41
EffNetv2-M	75.8 ± 0.47	60.4 ± 1.24	82.1 ± 0.57	69.8 ± 0.65
EffNetv2-M + CWT	76.9 ± 0.40	<b>61.0 ± 0.66</b>	82.5 ± 0.43	69.9 ± 0.64
EffNetv2-M + CWT + WD	76.5 ± 0.75	<b>61.0 ± 0.66</b>	82.5 ± 0.31	<b>70.2 ± 0.83</b>
EffNetv2-M + CWT + WD*	77.0 ± 0.46	60.3 ± 0.80	82.4 ± 0.38	69.2 ± 0.70
EffNetv2-M + CWT + AUG	78.1 ± 0.49	60.4 ± 0.24	80.2 ± 0.38	68.9 ± 0.51
EffNetv2-M + CWT + AUG + WD	78.0 ± 0.50	60.5 ± 0.42	80.3 ± 0.62	68.6 ± 0.24

Note: Different versions of the models include class weighted training (CWT), weight decay (WD), decoupled weight decay (WD\*), and heavy augmentation (AUG); see text for details. The model versions in boldface show progressive improvement and are also displayed in Figure 5. The best combinations for EffNet-b4 and EffNetv2-M (indicated by bold figures) employ different components. The displayed figures are averages and standard deviations over 5 independent runs (each run is a 5-fold cross-validation).

#### 4.4 | Experiments on model ensembles

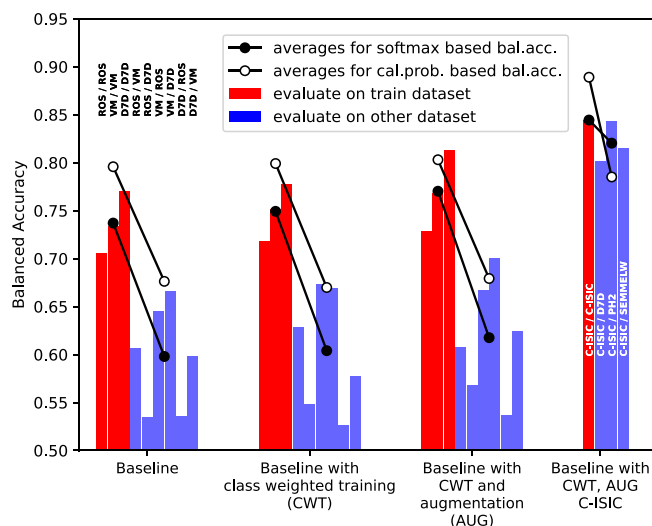
The classification performance can be improved by employing ensembles of models. In the experiments below, we used different combinations of the models described in Section 3.3. Since the three feature classifiers and the seven derm7pt classifiers provide complementary information to the 18 strong backbone classifier models; their results are best combined by a committee machine, which learns to take these properties into account – direct averaging may yield poor results. The models of the 22 ensembles are trained on the combined ISIC (C-ISIC) dataset, while the seven derm7pt classifiers are naturally trained on DERM7D.

The balanced accuracy values of these ensembles are shown in Table 4. The following observations can be made. The simple average of the softmax output of the 18 model ensemble (AVG-18), thresholded at 0.5, shows poor performance, comparable with the C-ISIC Baseline. The results are much better when the calibrated probabilities of the 18 models are averaged (Calib.Prop.AVG-18). When instead of averaging, a trained committee machine fuses the softmax outputs of the 18 models (Committee-18), the results are about the same as for Calib.Prop.AVG-18. Adding further models, first the four feature models (Committee-22) and then the remaining 7 derm7pt models (Committee-29) bring improvements in all datasets except for dataset PH2. For both the DERM7D and the Semmelweis University dataset, our

Committee-29 ensemble performed best. The dataset DERM7D has a special role: it is a training dataset for Committee-29 (for 7 of its models), while it is an “other” (non-training) dataset for all the other ensembles. Even when restricting ensembles for which DERM7D is non-training, still our model (Committee-22) is the best.

As we have seen, employing calibrated probabilities can be very beneficial (e.g., when predicting on a holdout subset of the training set, when training on small datasets, and when using models trained on an unbalanced dataset without class-weighted training). However, in other circumstances, it can worsen the results due to the fit to some special statistical features (via the optimization of the sigmoid function) of the training set since that sigmoid may differ from dataset to dataset. The latter was observed in the committee machine-based ensembles: the prediction benchmarks on other datasets were better when the committee machine used the softmax output of the convnets, not their calibrated probabilities.

So far we compared various model versions based on balanced accuracy, which is a suitable measure to use for unbalanced datasets. However, to help comparing with literature results, in Table 5 the sensitivity, specificity,  $F_1$  score, and raw accuracies are displayed for our best dermoscopic model, Committee-29. While direct comparison is difficult (as due to dataset dependence the choice of training and testing dataset is crucial), our model compares favorably to literature results. See the Appendix for a detailed comparison.



**FIGURE 5** Dataset dependence: the figure depicts the values of balanced accuracy achieved by the baseline model for different training and testing combinations. The models are trained and evaluated via cross-validation on the ROS and VM sub-datasets of HAM10000, on the dermoscopic subset (DERM7D) of the derm7pt dataset, and on the combined ISIC (C-ISIC) dataset. Red bars indicate the evaluation of the test sets on the respective training dataset. Blue bars show the evaluation of other datasets. Individual columns of the four-column groups are denoted as X/Y showing the training (X) and the testing set (Y), respectively. For example, VM/ROS denotes that training happened on VM, and evaluation was done on the ROS dataset. Out of the four main column groups, the first three groups are as follows: Baseline, Baseline with class-weighted training (CWT), and Baseline with class-weighted training and heavy augmentation (AUG). The last group (C-ISIC) uses the same model as AUG but was (A) trained on the combined ISIC dataset, and (B) evaluated on DERM7D and on two small datasets: ROS and VM were left out from these comparisons since they are part of the ISIC dataset. For each model type, the balanced accuracy averaged over the three training datasets, and over the other datasets (six combinations) are marked with solid black disks. Balanced accuracy is also calculated via thresholding at 0.5 the “calibrated probabilities.” For this case, only the averages (of the red and blue values) are displayed as open circles.

#### 4.5 | Experiments on macro images without transfer learning

To test our ensemble for classification from clinical images, first, we performed a set of experiments, in which the deep convnets were trained only on dermoscopic images. Then the macro images were predicted on these dermoscopic-trained deepnets, and the outputs were fed to train the committee machine. Results for balanced accuracies are shown in Table 6.

We observed that the ensemble performed reasonably well on macro images even without fine-tuning the deep convnets on macro images. Interestingly training on the

very small MED-NODE dataset yielded comparable results with the medium-size DERM7M and PAD-UFES datasets.

#### 4.6 | Experiments on macro images with transfer learning

Finally, we performed experiments when some of the deep convnets have been fine-tuned on macro images. As before, we trained our 11 deep convnets (with the usual ImageNet weight initialization) on the C-ISIC and the DERM7D dermoscopic datasets, and employed the 18 models trained by.<sup>16</sup> After obtaining the models, we examined the effect of transfer learning (fine tuning) with our 11 models on the macro image set of the Derm7pt dataset, while keeping the resource-hungry 18 models fixed. We also experimented with SAM<sup>12</sup> and its adaptive version ASAM<sup>22</sup> during the cross-training step, to reduce over-fitting to the training database. We predicted with the trained ensembles on the macroscopic datasets - DERM7M, MED-NODE, and PAD-UFES - to obtain softmax values for training and evaluating the committee machine. The results are shown in Table 7.

It is interesting to note that for traditional (no SAM) transfer learning the fine-tuning of the deepnets (3rd line in Table 7) produced worse 3-dataset average results, than no transfer learning (first and second lines). This can be a sign of dataset-level overtraining: the small sample size of the macro training set could not prepare the convnets to the spectrum of samples from other macro datasets. This argument is underpinned by the fact that the cross-validation results for the training DERM7M dataset increased for deepnet refinement (from about 70% to 72%), as the deepnets became more specialized for that particular dataset.

We have overcome this problem with adaptive SAM regularization that achieved the best results when the regularization affected both the deep convnets and the committee machine.

### 5 | DISCUSSION AND CONCLUSIONS

There are a number of reasons why skin lesion datasets differ from each other. According to a recent taxonomy,<sup>35</sup> a dataset can be “biased” from an idealized dataset due to technical reasons like device bias (e.g., resolution or sharpness of images, black lines near edges, device artifacts like millimeter scales) or capture bias (light conditions, image cropping policy resulting in different lesion area fraction distribution).

**TABLE 2** Performance of the EffNet-b4 + CWT + AUG version of the model for transformed image sets. The VM, ROS, and DERM7D image sets were altered to have equal lesion-to-image size distribution, similar image edges (removal of artifacts), and equalized color component distributions. The displayed figures are averages and standard deviations over five independent runs

Equalizing	Balanced accuracy		Bal. Acc. on Cal. Pr.	
	Train	Other	Train	Other
Nothing (original)	77.0 ± 0.39	61.8 ± 0.22	80.3 ± 0.68	67.9 ± 0.56
Area	76.6 ± 0.50	62.3 ± 0.18	80.8 ± 0.86	68.3 ± 0.33
Area + edges	77.0 ± 0.17	<b>62.5</b> ± 0.33	80.7 ± 0.62	<b>68.7</b> ± 0.26
Area + edges + color	76.7 ± 0.26	61.5 ± 0.55	80.6 ± 0.48	68.0 ± 0.36

Note: The best values obtained on Other datasets are indicated by bold text.

**TABLE 3** C-ISIC training: Balanced accuracy, and balance accuracy evaluated on the calibrated probabilities of the EffNet-b4 + CWT + AUG model trained on the C-ISIC dataset

Dataset	Balanced accuracy	Bal. Acc. on Cal. Pr.
Train: C-ISIC	84.4 ± 0.27	88.9 ± 0.17
Other: DERM7D	80.2 ± 0.43	77.3 ± 0.54
Other: PH2	84.4 ± 1.06	75.9 ± 0.80
Other: Semmelweis	81.5	82.3

Note: The data is also plotted as the last group in Figure 5. The displayed figures are averages and standard deviations over five independent runs, except for Semmelweis where no error is calculated due to limited data availability.

We compensated for the technical differences and found that for three mono-source datasets that such technical deviations accounted for only 5% of the dataset dependence. It then follows that the rest is due to more subtle differences. Excluding technical differences, database dependence can be due to sampling bias, such as ethnic differences that affect (a) background skin color and also (b) the distribution of anatomic sites, both influencing the visual appearances of the lesions. Beyond that, the varying amount of UV radiation in different geographic regions affects (c) tumor characteristics.<sup>44</sup>

We have studied the causes of variances in the evaluations for different databases. We found that the classification of dermoscopic skin lesion images has high dataset dependence: the balanced accuracy drops by 14%–15% for the medium-size mono-source datasets we tested. Standard machine learning techniques aimed at improving the predictive power of the models, including class, weighted averaging (in case of imbalanced datasets), heavy data augmentation, regularization by weight decay, and the use of calibrated probability provide some incremental benefit, but the improvement is small, and the dataset dependence gap remains roughly unchanged. Combination of compensation, for example, class-

weighted averaging, and regularization, such as weight decay and heavy augmentation seems worthwhile in our cases, but with a caveat: the winning method depends on the classifier backbone network.

The first real improvement is brought about by training the models on large, *multi-source* datasets (see Table 3). This way, the dataset dependence is reduced to a few percent. The measured balanced accuracy is excellent for PH2 (no dataset dependence compared to the training C-ISIC dataset), and also good for DERM7D and for Semmelweis. The calibrated probabilities, however, in the case of multi-source training give rise to higher dataset dependencies. Presumably, the extra fit to the distribution of the large training dataset differs from the distribution of the testing dataset which increases database dependence somewhat. If the distribution of the dataset is known then the method of calibrated probabilities could be applied.

The second relevant improvement is to employ a diverse ensemble of models. Our ensemble consisted of 29 models including powerful general-purpose classifiers, models fed with preprocessed input to focus on important features, and models directly trained on specific features instead of the final target. The output of the constituent models was fused by a trained method: a shallow net committee machine, as conventional aggregation methods like averaging or majority voting is unsuitable for models with diverse training target. This ensemble performed well both on the training datasets and on the other datasets (Table 4). The dataset dependence, which was 14%–15% for single models trained on mono-source datasets, is reduced to a few percent for the ensemble model.

We also aimed at classifying clinical (macro) images of skin lesions, which is a harder problem than dermoscopic image classification. We observed that the ensemble performed reasonably well on macro images even without fine-tuning the deep convnets on macro images when the macro image training was conducted on the level of the committee machine (Table 6).

**TABLE 4** Values of balanced accuracy evaluated for a number of ensemble models. “Out-of-fold”: during cross-validation predict the hold-out set for the given fold, then average the metrics over all five folds

Balanced accuracy (%)	Predict on out-of-fold		Predict on other datasets		
	C-ISIC	DERM7D	DERM7D	PH2	Semmelweis
AVG-18	89.12		79.2	84.1	77.0
Calib.Prob.AVG-18	93.45		82.5	91.9	81.8
Committee-18	93.44		83.6	90.9	83.2
Committee-22	93.56		84.8	90.9	83.1
Committee-29	93.51	84.8		91.2	83.6
Estimated error	$\pm 0.04$	$\pm 0.2$	$\pm 0.2$	$\pm 0.3$	$\pm 0.4$

*Note:* The ensemble models are detailed in the main text. All models except Committee-29 are trained on the combined ISIC dataset only. For Committee-29, 22 of its constituent models are trained on the combined ISIC dataset and the remaining 7 models are in the DERM7D dataset. All models are evaluated on the combined ISIC dataset, the dermoscopic subset of the derm7pt dataset (DERM7D), the PH2 dataset, and Semmelweis University's dataset.

**TABLE 5** Various performance metrics for the Committee-29 model, in addition to the balanced accuracy presented in Table 4

Committee-29	Predict on out-of-fold		Predict on other datasets	
	C-ISIC	DERM7D	PH2	Semmelweis
Sensitivity	92.6	84.5	90.0	86.3
Specificity	94.4	85.2	92.5	81.0
$F_1$ score	73.8	73.8	81.8	76.5
Accuracy	94.2	85.1	92.0	82.7

**TABLE 6** Balanced accuracy of the best committee machines trained on each dataset and evaluated on all three datasets

Train dataset	Test dataset			Test dataset
	DERM7M	PAD-UFES	MED-NODE	Average
DERM7M	$69.7 \pm 0.7$	$71.2 \pm 1.1$	$89.4 \pm 1.1$	$76.8 \pm 0.6$
PAD-UFES	$68.1 \pm 0.8$	$78.5 \pm 1.4$	$87.7 \pm 1.1$	$78.1 \pm 0.7$
MED-NODE	$73.6 \pm 1.6$	$72.8 \pm 0.8$	$85.1 \pm 2.0$	$77.2 \pm 0.9$

*Note:* The last column displays the average over all three datasets. No transfer learning has been performed on the deepnets, the committee machine was trained with SAM.

**TABLE 7** Balanced Accuracy of different versions of the model ensemble trained on the DERM7M dataset, and evaluated on the training dataset (by means of cross-validation) and three other macro datasets

Transf. learning in deepnets	Committee machine	DERM7M (train + test)	PAD-UFES (test)	MED-NODE (test)	Average over 3 datasets
No	Plain	$70.2 \pm 1.1$	$71.6 \pm 0.7$	<b><math>89.4 \pm 0.9</math></b>	$77.1 \pm 0.5$
No	SAM	$69.7 \pm 1.2$	$71.2 \pm 1.1$	<b><math>89.4 \pm 1.1</math></b>	$76.8 \pm 0.7$
Plain	Plain	$72.0 \pm 0.6$	$69.2 \pm 1.0$	$81.4 \pm 1.1$	$74.2 \pm 0.5$
Plain	ASAM	$73.2 \pm 0.6$	$73.1 \pm 1.1$	$87.5 \pm 1.3$	$77.9 \pm 0.6$
ASAM	Plain	$74.4 \pm 1.5$	$72.4 \pm 0.4$	$86.6 \pm 1.2$	$77.8 \pm 0.7$
ASAM	ASAM	<b><math>74.9 \pm 1.1</math></b>	<b><math>74.1 \pm 0.5</math></b>	$88.7 \pm 1.0$	<b><math>79.2 \pm 0.5</math></b>

*Note:* The last column displays the average over the three datasets. For table entries on SAM/ASAM training, only the better of the two is displayed to reduce clutter (which was ASAM with one exception). The best values for each dataset are indicated by bold text.

With the aim for further improvements, we fine-tuned 11 of the convnet models of the ensemble on macro images from the DERM7M dataset. With

traditional training, that is, without (adaptive) SAM, cross-validation results on the training DERM7M dataset increased, but surprisingly on the test datasets,



deterioration has been observed. According to our results, regularization using adaptive SAM for both the deepnets and the committee machine is the superior strategy the best strategy for improving the predictive power of databases with unknown or imprecise statistical properties that could be taken into account using the probability calibration procedure (Table 7).

## ACKNOWLEDGMENTS

This research was funded by the “Application Domain Specific Highly Reliable IT Solutions” project, implemented with the support provided by the National Research, Development and Innovation Fund of Hungary financed under the Thematic Excellence Program no. TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme; the Hungarian Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program; and the European Union co-financed by the European Social Fund (EFOP-3.6.3-16-2017-00002).








## CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## DATA AVAILABILITY STATEMENT

Code and data accompanying this paper are available at: <https://github.com/semiquark1/skin>.

## ORCID

Ellák Somfai  <https://orcid.org/0000-0002-2218-8855>  
 Benjámin Baffy  <https://orcid.org/0000-0002-1775-180X>  
 Kristian Fenech  <https://orcid.org/0000-0002-8288-9303>  
 Dorina Korózs  <https://orcid.org/0000-0002-1445-8739>  
 Marcell Pólik  <https://orcid.org/0000-0002-0069-8379>  
 Miklós Sárdy  <https://orcid.org/0000-0003-4306-5093>  
 András Lőrincz  <https://orcid.org/0000-0002-1280-3447>

## REFERENCES

- Ahmad B, Usama M, Ahmad T, Khatoon S, Alam CM. An ensemble model of convolution and recurrent neural network for skin disease classification. *Int J Imaging Syst Technol*. 2022; 32(1):218-229.
- Benyahia S, Meftah B, Lézoray O. Hierarchical approach for the classification of multi-class skin lesions based on deep convolutional neural networks. *International Conference on Pattern Recognition and Artificial Intelligence*. Springer; 2022: 139-149.
- Brinker TJ, Hekler A, Enk AH, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur J Cancer*. 2019; 113:47-54.
- Brinker TJ, Hekler A, Utikal JS, et al. Skin cancer classification using convolutional neural networks: systematic review. *J Med Internet Res*. 2018;20(10):e11936.
- Bröcker J, Smith LA. Increasing the reliability of reliability diagrams. *Weather Forecast*. 2007;22(3):651-661.
- Codella N, Cai J, Abedini M, Garnavi R, Halpern A, Smith JR. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. *International Workshop on Machine Learning in Medical Imaging*. Springer; 2015:118-126.
- Codella N, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*. 2019.
- Codella NC, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE; 2018:168-172.
- Combalia M, Codella NC, Rotemberg V, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*. 2019.
- Cormier JN, Xing Y, Ding M, et al. Ethnic differences among patients with cutaneous melanoma. *Arch Intern Med*. 2006; 166(17):1907-1914.
- Filali Y, El Khoukhi H, Sabri MA, Aarab A. Analysis and classification of skin cancer based on deep learning approach. *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*. IEEE; 2022:1-6.
- Foret P, Kleiner A, Mobahi H, Neyshabur B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*. 2020.
- Gessert N, Nielsen M, Shaikh M, Werner R, Schlaefel A. Skin lesion classification using ensembles of multi-resolution efficient nets with metadata. *MethodsX*. 2020;7:100864.
- Ghiasvand R, Robsahm TE, Green AC, et al. Association of phenotypic characteristics and UV radiation exposure with risk of melanoma on different body sites. *JAMA Dermatol*. 2019; 155(1):39-49.
- Giotis I, Molders N, Land S, Biehl M, Jonkman MF, Petkov N. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Syst Appl*. 2015;42(19): 6578-6585.
- Ha Q, Liu B, Liu F. Identifying melanoma images using efficient net ensemble: winning solution to the siim-isic melanoma classification challenge. *arXiv preprint arXiv:2010.05351*. 2020.
- Jojoa Acosta MF, Caballero Tovar LY, Garcia-Zapirain MB, Percybrooks WS. Melanoma diagnosis using deep learning techniques on dermoscopic images. *BMC Med Imaging*. 2021; 21(1):1-11.
- Jolliffe IT, Stephenson DB. *Forecast Verification: a practitioner's Guide in Atmospheric Science*. John Wiley & Sons; 2012.
- Kawahara J, BenTaieb A, Hamarneh G. Deep features to classify skin lesions. *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2016:1397-1400.
- Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J Biomed Health Inform*. 2019; 23(2):538-546.
- Kousis I, Perikos I, Hatzilygeroudis I, Virvou M. Deep learning methods for accurate skin cancer recognition and mobile application. *Electronics*. 2022;11(9):1294.
- Kwon J, Kim J, Park H, Choi IK. ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural

- networks. *International Conference on Machine Learning*. PMLR; 2021:5905-5914.
23. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26(6):900-908.
  24. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. 2017.
  25. Mendonça T, Ferreira PM, Marques JS, Marcal AR, Rozeira J. PH2 - a dermoscopic image database for research and benchmarking. *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2013:5437-5440.
  26. Menzies SW. *An Atlas of Surface Microscopy of Pigmented Skin Lesions: Dermoscopy*. McGraw Hill Professional; 2003.
  27. Menzies SW, Ingvar C, Crotty KA, McCarthy WH. Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. *Arch Dermatol*. 1996; 132(10):1178-1182.
  28. Mohan SV, Chang ALS. Advanced basal cell carcinoma: epidemiology and therapeutic innovations. *Curr Dermatol Rep*. 2014; 3(1):40-45.
  29. Murphy AH, Winkler RL. Reliability of subjective probability forecasts of precipitation and temperature. *J R Stat Soc Ser C Appl Stat*. 1977;26(1):41-47.
  30. Naeem A, Farooq MS, Khelifi A, Abid A. Malignant melanoma classification using deep learning: datasets, performance measurements, challenges and opportunities. *IEEE Access*. 2020;8: 110575-110597.
  31. Nasr-Esfahani E, Samavi S, Karimi N, et al. Melanoma detection by analysis of clinical images using convolutional neural network. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2016:1373-1376.
  32. Rastgoo M, Garcia R, Morel O, Marzani F. Automatic differentiation of melanoma from dysplastic nevi. *Comput Med Imaging Graph*. 2015;43:44-52.
  33. Rotemberg V, Kurtansky N, Betz-Stablein B, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci Data*. 2021;8(1):1-8.
  34. Somfai E, Baffy B, Fenech K, et al. Minimizing false negative rate in melanoma detection and providing insight into the causes of classification. *arXiv preprint arXiv:2102.09199*. 2021.
  35. Srinivasan R, Chander A. Biases in AI systems. *Commun ACM*. 2021;64(8):44-49.
  36. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R, eds. *Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research*. PMLR; 2019:6105-6114.
  37. Tan M, Le QV. Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning*. PMLR; 2021:10096-10106.
  38. Tresp V. Committee machines. *Handbook for Neural Network Signal Processing*. CRC Press; 2001:1-18.
  39. Tsatmali M, Ancans J, Thody AJ. Melanocyte function and its control by melanocortin peptides. *J Histochem Cytochem*. 2002; 50(2):125-133.
  40. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5:180161.
  41. Venkateswararao N, Rao PV. Distributed densely connected convolutional network approach on patient's metadata of dermoscopic images for early melanoma detection. *Int J Health Sci*. 2022;6(2):5446-5456.
  42. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. Eca-net: efficient channel attention for deep convolutional neural networks. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2020:11534-11542, virtual.
  43. Yang D, Salciccioli J, Marshall D, Sheri A, Shalhoub J. Trends in malignant melanoma mortality in 31 countries from 1985 to 2015. *Br J Dermatol*. 2020;183:1056-1064.
  44. Yardman-Frank JM, Glassheim E, Kricker A, et al. Differences in melanoma between Canada and new south Wales, Australia: a population-based genes, environment, and melanoma (gem) study. *JID Innov*. 2021;1(1):100002.
  45. Yu K-H, Lee T-LM, Yen M-H, et al. Reproducible machine learning methods for lung cancer detection using computed tomography images: algorithm development and validation. *J Med Internet Res*. 2020;22(8):e16709.
  46. Yu L, Chen H, Dou Q, Qin J, Heng P-A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans Med Imaging*. 2016;36(4):994-1004.
  47. Zhang J, Xie Y, Xia Y, Shen C. Attention residual learning for skin lesion classification. *IEEE Trans Med Imaging*. 2019;38(9): 2092-2103.

**How to cite this article:** Somfai E, Baffy B, Fenech K, et al. Handling dataset dependence with model ensembles for skin lesion classification from dermoscopic and clinical images. *Int J Imaging Syst Technol*. 2023;33(2):556-571. doi:10.1002/ima.22827

## APPENDIX A: CALIBRATED PROBABILITY

We obtain calibrated probability as follows. For a given dataset selected for calibration (typically the training dataset), the softmax predictions of the classifier are collected. In the case of an imbalanced dataset, the minority class is upsampled to obtain a balanced set of samples. Ideally, the distribution of softmax values is roughly uniform, so a histogram of equal bin sizes would contain comparable samples per bin. If this is not the case, for example, the values tend to concentrate around 0 and 1, they are transformed by the composition of an inverse sigmoid and a sigmoid function to increase uniformity:

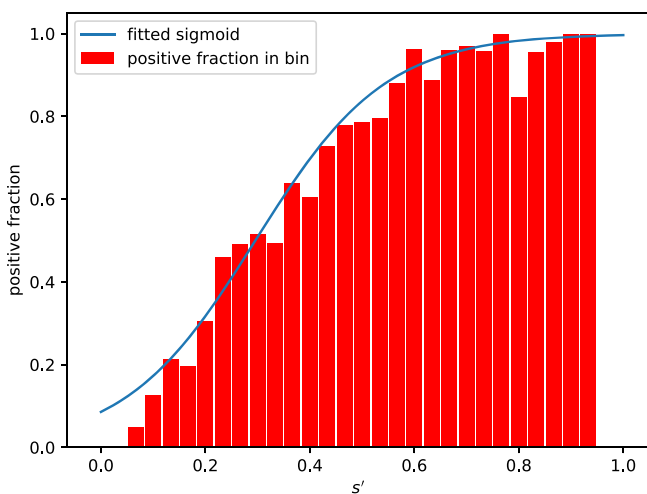
$$s' = \sigma_{\beta,0}(\sigma_{1,0}^{-1}(s)),$$

where the sigmoid function is parametrized as

$$\sigma_{\beta,x_0}(x) = \frac{1}{1 + \exp - \beta(x - x_0)},$$

and the superscript  $-1$  denotes the inverse of the function. The ground truth 0 or 1 values of the samples are then collected into  $n$  uniformly separated bins in the unit interval according to the modified prediction  $s'$ . The average within a bin corresponds to the fraction of positive samples with similar  $s'$  values. This is expected to cross over from 0 to 1 as  $s'$  increases: a sigmoid function (with two parameters) is fit on the empirical observations, see Figure A1 below for illustration.

At prediction time, the softmax value  $s$  is outputted by the classifier, which is converted to  $s'$  using



**FIGURE A1** Probability calibration curve for the Baseline model trained and calibrated on the ROS dataset

parameters fixed for the training set. Then the calibrated probability is the fitted sigmoid function (again, fixed at training time) evaluated at  $s'$ . The predicted probability calculated this way is a strictly increasing function of  $s$ .

## APPENDIX B: TECHNICAL DETAILS FOR THE SINGLE MODELS

The EfficientNet-b4 models were trained on cross-entropy loss for 15 epochs, and the learning rate followed cosine annealing with initial ramp-up and maximal value of  $3 \times 10^{-4}$ . The images were rescaled to the nominal image size  $380 \times 380$  of EfficientNet b4.

To improve performance, heavy augmentation was used, including flips, axis transpose, random shift, scale changes, and rotations; brightness and contrast adjustments, image color adjustment in hue, saturation, and value; various blur methods (motion, median, Gaussian); optical, grid and elastic distortions; pixel level (Gaussian) noise and block erasure as well as the CLAHE (contrast limited adaptive histogram equalization) method.

The EfficientNetv2-M-based classifier was similar to the EfficientNet-b4-based one, with the exception of using an image size of  $480 \times 480$  pixels. The maximum value of the learning rate schedule was  $2 \times 10^{-4}$  (optimized by grid search).

## APPENDIX C: EQUALIZING GLOBAL DATASET IMAGE PROPERTIES

As Figures 3 and 4 demonstrate, the statistical properties of the ROS, VM, and DERM7D datasets are not identical. We remedy this by the following procedure. To equalize the fraction histograms of the lesion areas, we cropped or padded the images to achieve identical cumulative distribution functions. If  $\text{cdf}_X(a) = \Pr[\text{area fraction} < a]$  is the cumulative distribution function of the lesion area to image area fraction in dataset  $X$ , then for example for a given image in dataset VM with area fraction  $a$ , the image is cropped or padded such that the area fraction becomes  $a' = \text{cdf}_{ROS}^{-1}(\text{cdf}_{VM}(a))$  where superscript  $-1$  denotes inverse. The black lines at the edges of some DERM7D images were removed by appropriate cropping. Finally, the color component histograms were equalized: for hue, a constant periodic shift, whereas for saturation and value suitable strictly increasing piece-wise linear transfer functions were applied.

## APPENDIX D: COMPARISON WITH RECENT WORKS IN THE LITERATURE

It is natural to assume that the ultimate goal of melanoma classification is to evaluate a skin lesion image coming from an unknown source. As we have shown in the main paper, due to dataset dependence that this is a more difficult task than evaluation on a hold-out test set of a given (training) dataset. Therefore it is appropriate to present the performance of our best dermoscopic model (Committee-29) as in Table 4, where in addition to the out-of-fold training set performance (93.5% balanced accuracy for C-ISIC, and 84.8% for DERM7D which was involved in training in 7 out of the 29 constituent models), values for the disjunct datasets PH2 (91.2%) and Semmelweis (83.6%) are also presented. The clinics where PH2 and Semmelweis images were taken are completely independent of the clinics involved in the ISIC datasets and DERM7. There was a strong reason to select the given training and testing datasets (the heterogenous multi-dataset nature of C-ISIC together with DERM7D, and the independent clinic source of PH2 and Semmelweis). Due to dataset dependence, these figures cannot be directly compared with literature results, where the choice of training and testing dataset is not identical to ours.

One state-of-the-art result, with which we *can* compare directly, is the winner of the 2020 SIIM-ISIC Melanoma Classification Challenge,<sup>16</sup> see “AVG-18” entry in Table 4. For fairness, that model was developed with a ROC AUC metric in mind, so for our metrics, the balanced accuracy (with decision threshold at 0.5 level), using calibrated probability is more reasonable, see “Calib.Prob.AVG-18” entry. Compared to that *improved* model, our Committee-29 is no worse on C-ISIC and PH2 (differences are within error) and is better on the Semmelweis dataset. (Also better on the DERM7D dataset, but for Committee-29 that one became a training dataset.)

For other recent results since 2021, only rough comparisons can be made, since the training and testing

datasets (typically the same dataset there) are different from our choices. Nevertheless neither of the five works cited below reach the 93.5% balanced accuracy value that we obtain for C-ISIC, even though we consider the cross-dataset predictions more important.

While the multi-step approach of,<sup>17</sup> with 87.2% balanced accuracy on ISIC-2017 (training + testing) is better than our single-model single-dataset figures (80% for EffNet-b4 and 82% for EffNetv2-M, trained and tested on smaller datasets than ISIC-2017); our Committee-29 is significantly better on C-ISIC (93.5%), which contains ISIC-2017, although has much larger sample size. Reference 41 does not clearly state which version of the ISIC dataset is used for training and testing and does not cite balanced accuracy or sensitivity and specificity figures (their highest accuracy value is 82%, but for unbalanced datasets typical for skin lesions raw accuracy is misleading). The hierarchical approach of<sup>2</sup> (their highest balanced accuracy is 86.2%, interestingly for the smaller backbone tried, and without augmentation) yields lower figures for ISIC-2019, but this is for 8-way classification, not binary as ours. The results of<sup>11</sup> are worse than ours (at most 66.5% balanced accuracy for PH2, and 55.5% for ISIC2017) for all tested backbones. Kousis et al.<sup>21</sup> considered both 7-way and binary classification of skin lesion images (for binary the classes were benign vs. malignant, with malignant containing non-melanoma cancer types as well). Their best binary benchmark (89% balanced accuracy for the full HAM10000) is better than our single model results (on subsets of the HAM10000 datasets) but worse than Committee-29 on C-ISIC. It may be worth pointing out that some of the lesions in HAM10000 have multiple images (though not identical), so a naive train-test split might be liable for a small amount of data leakage. We properly deal with this issue in our mono-source experiments by keeping only a single image per lesion in the sub-datasets of HAM10000 (but we can not for C-ISIC, as lesion id is not provided; however we expect the effect is smaller there due to the larger number of independent data sources).