Superconverging Deep Learning Model for Skin Lesion Classification

David Dueñas Gaviria
MSc.
Facultat d'informatica de Barcelona
Universitat Politècnica de Catalunya
Spain
Email: blobquiet@gmail.com

Dr, Md Mostafa Kamal Saker[*]
Senior Research Associate
PhD.
Department of Engineering Science
University of Oxford
United Kingdom
m.sarker@eng.ox.ac.uk

Dr. Petia Ivanona Radeva[†]
PhD.
Department of Mathematics
and Computer Science
Universitat de Barcelona
Spain
petia.ivanova@ub.edu

The field of computer vision has for years been dominated by Convolutional Neural Networks (CNNs) in the medical field. However, there are various other Deep Learning (DL) techniques that have become very popular in this space. Vision Transformers (ViTs) are an example of a deep learning technique that has been gaining in popularity in recent years. In this work, we study the performance of ViTs and CNNs on skin lesions classification tasks, specifically melanoma diagnosis. We compare the performance of ViTs to that of CNNs and show that regardless of the performance of both architectures, an ensemble of the two can improve generalization. We also present an adaptation to the Gram-OOD method (detecting Out-of-distribution (OOD) using Gram matrices) for skin lesion images. A rescaling method was also used to address the imbalanced dataset problem, which is generally inherent in medical images. The phenomenon of super-convergence was critical to our success in building models with computing and training time constraints. Finally, we train and evaluate an ensemble of ViTs and CNNs, demonstrating that generalization is enhanced by placing first in the 2019 and third in the 2022 ISIC Challenge Live. Leaderboard (available at https://challenge.isic-archive.com/leaderboards/live/).*

Nomenclature

NNs	Neural Networks
CNNs	Convolutional Neural Networks
ViTs	Vision Transformers
MCM	Mean Correlation Matrix

1 Introduction

Skin cancer has become a major public health concern, between 2 and 3 million non-melanoma skin cancers occur each year and 132 thousand melanoma worldwide, claiming more than 20 thousand lives in Europe alone each year, and 57 thousand worldwide, based on the most recent [1], [2]. According to a study by [3] from the International Agency for Research on Cancer (IARC), "*the number of new cases of cutaneous melanoma per year will increase by more than 50% from 2020 to 2040*", implying that the burden of melanoma will only increase in the future as the population ages. Likewise, melanoma is the deadliest form of skin cancer [1], and a later stage of melanoma diagnosis has been linked to a significant increase in mortality rate.

Ultraviolet (UV) radiation from the sunlight, which we are all exposed to on a daily basis, has been identified as the primary environmental risk factor for the development of melanoma skin cancer [4], and yet, within a melanoma diagnosis, the 5-year survival rate exceeds 90% [5]; this final is a major motivation for research efforts unfolding worldwide to shift its diagnosis toward earlier stages, to prevent its occurrence, and allow the development of earlier treatments. A study on the feasibility of applying deep learning methods to address this issue is promoted here to classify skin lesions and evaluate them through the use of general-purpose neural architectures focused on improving its classification perfor-

[*] Co-supervisor.
[†] Supervisor.

mance and assessing it particularly on the melanoma.

As medical professionals' and patients' needs for technology have increased, so have the demands for automated skin cancer diagnosis [6]; In response, current research has produced automated skin cancer diagnostic tools that perform on par with dermatologists who rely mostly on visual diagnosis, dermoscopic analysis, or invasive biopsy, alone with a histopatological study. Likewise, Deep Learning (DL) has revolutionized the field of computer vision in recent years with the resurgence of Neural Network (NNs) architectures [7]. Convolutional Neural Networks (CNNs) have become the dominant DL technique in this field, due in large part to their success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [8]. However, there are a number of other DL techniques that have been gaining in popularity in recent years. Particularly Vision Transformers (ViTs) [9], which correspond to a type of transformer that is specifically designed for computer vision tasks. Transformers are a type of DL model that are based on the attention mechanism and have proved successful in a number of natural language processing tasks [10]. Although considerable research has been done on the use of ViTs for medical image classification, see [11], robustness againts skin lesions in generalization has not yet being explicit. This is generally the case because the training and testing data for many closed-world tasks are taken from the same distribution. However, in the ISIC 2019 dataset particuarly, the effect of an outlier class poses a significant challenge for ViTs in comparison to traditional CNNs. Hence, the aim with this study is to answer: How useful is the incorporation of ViTs in classification for skin cancer detection, particularly melanoma, in comparison to CNNs?

A cooperation between academics and business called the International Skin Imaging Collaboration (ISIC) aims to make it easier to use digital skin imaging to lower the death rate from melanoma. ISIC engages the dermatological and computer science communities in the creation and promotion of standards for digital skin imaging in order to advance diagnosis. In addition to directly assisting in the diagnosis of melanoma through tele-dermatology, clinical decision support, and automated diagnosis, digital images of skin lesions are being used to educate professionals and the general public on the recognition of melanoma.

Skin lesion classification using ViTs and CNNs shares the same goal of detecting disease lesions by using image-level and patient-level data. Thus, it makes sense to test their performance together using a common ensemble. To give a brief, the contributions of this study are as follows:

1. Focusing on the main goal of skin disease classification problem, we propose a robust model outperforming the state of the art in 2019 ISIC competition, based on an ensemble that comprises a wide range of model architectures, including top accuracy ViTs and popular CNNs.
2. We provide a consistent validation pipeline supported on the widely studied super-convergence phenomenon which allowed for a larger number of individual experiments, despite computing and time constraints.
3. Our model shows improvements on the Gram-OOD* method for the detection of Out-of-distribution samples in the ensemble predictions.
4. Instead of penalizing a loss function in training, our model demonstrated that the inherent inductive bias of skin lesion diagnosis due to the imbalanced data can be handled by rescaling the decision threshold at model inference.
5. Finally we demonstrated that preserving semantic-transformations is crucial in a data augmentation regime achieving top performance with our combined ViTs and CNNs ensemble model.

The submission rankings in this study reached first place in the ISIC 2019 live challenge with a balanced multi-class accuracy (BACC) of 0.670, and top ten for melanoma classification in the ISIC-2020 live challenge with an Area Under the Curve (AUC) score of 0.940. We used the same target prediction for the malignant melanoma, indicating strong generalization potential to close the gap in considering deep learning techniques as a reliable source for an early diagnosis.

The following study is arranged as follows: the next chapter focuses on the data, including who hosts it and what it consists of. The third chapter goes over our model description and implementation processes training the various ViTs and CNNs models used and generating predictions. The fourth chapter displays and summarizes all of the results acquired along with the discussion on the validation approach before providing predictions. On the whole, the last gives a conclusion of the study given and future research lines to be pursued.

2 Skin lesion datasets

At the image level, there are 9.1 GB worth 25,331 dermoscopic images available for training in 8 different classes. This information was obtained from the Memorial Sloan Kettering Cancer Center, the BCN_20000 dataset from the Department of Dermatology, Hospital Clnic de Barcelona [12] and the HAM10000 dataset from the Department of Dermatology, Medical University of Vienna [13]. Table 1 shows the nine classes used for the diagnosis in this challenge. Likewise, the test dataset comprised 8,239 images with the extra outlier class that was not represented in the training data. Aside from the images, the collection includes metadata such as the patient's age and sex as well as the location of the individual skin lesion.

ISIC 2020 dataset [14] is composed of 23 GB worth 33,126 images of different resolutions for training and 10982 for the test set. A total of 2056 patients was gathered for this dataset at various locations around the world, including the Memorial Sloan Kettering Cancer Center in New York, the Melanoma Institute Australia and the Melanoma Diagnosis Centre in Sydney, the University of Queensland in Brisbane, the Medical University of Vienna, and the Hospital Clinic de Barcelona [14]. In contrast to the 2019 dataset, the unknown

Diagnosis	2019 dataset	2020 dataset	External data	
NV	12875 (50%)	5193 (15%)	7 point	1011
MEL	4522 (18%)	584 (2%)	PH2	200
BKL	2624 (10%)	223 (1%)	MED-NODE	170
UNK	0 (0%)	27126 (82%)	SD-198	5944
BCC	3323 (13%)		SKINL2V1-2-3	299
AK	867 (4%)			
SCC	628 (3%)			
VASC	253 (1%)			
DF	239 (1%)			
Total	25331	33126	7624	

Table 1. Diagnosis distribution for the 2019, 2020 ISIC datasets and the external dataset.

class accounted for the majority of benign occurrences, with Cafe-au-lait macule and atypical melanocytic proliferation, whereas the other three: melanocytic nevus, melanoma, and benign keratosis, are also shared diagnosis with the 2019 dataset.

With the presence of an outlier class in the ISIC 2019 dataset, it was reasonable to experiment with external data to attempt to increase training diversity for the unknown and generalization of the remaining classes. As an outline of [15], the outlier class for training was addressed through the usage of a subset of a collection of datasets, which are detailed below:

3 Method

Here we introduce our new method for skin lesion classification, which was able to demonstrate robustness in generalization by scoring first in the 2019 ISIC Challenge and third in the 2020 ISIC Challenge, despite computing and trainig-time limitations. Overall, the following contributions made it possible to achieve such a position:

1. Diversity provided by ViTs and CNNs ensemble.
2. Super-convergence, through the usage of the OneCycle LR in conjunction with the AdamP optimizer.
3. OOD detection through the usage of the Gram-OOD method.
4. Handling the imbalanced data problem, through rescaling the model' predictions, by using the output class probabilities.
5. Contextual image augmentation, for learning credible representations on the skin images.

3.1 Our Ensemble for Skin Lesion Classification

A variety of state-of-the-art ViTs and CNNs were explored in our work in order to study their jointly behaviour in the context of skin lesion diagnosis. After a thorough analysis on the state-of-the-art DL models and in particular those that made the top rank for 2019 live leaderboard, we concluded that the highly complex problem of skin lesion classification requires an ensemble of robust performing models. Hence, here we propose an ensemble that consists of:

(1) Data-efficient Image Transformer (DeiT) [16], which is a type of ViT trained using a teacher-student strategy specific to transformers relying on a distillation token ensuring that the student learns from the teacher through attention.

(2) EfficientNets [17], trained on Noisy-Student weights [18] and using a scaling technique to equally scale the network's width, depth, and resolution using a set of predefined scaling coefficients.

(3) ConvNeXt [19], resulting in a hybrid model lacking attention-based modules that adapt a ConvNet towards the design of a hierarchical Swin transformer.

The diagram of the pipeline is depicted in Figure 1, which shows the use of both ViTs and CNNs. Thus, the final ensemble in the training pipeline (a) shows in green and blue the ViTs and CNNs respectively, being trained using the 2019 ISIC dataset with additional external datasets (see Figure 6), with these considered as external data. The yellow line, on the other hand, represents the pipeline that was used to train the 2019 and 2020 ISIC datasets on images and metadata. (b), on the other hand, indicates the testing pipeline, which consisted of generating test predictions using TTA with a similar augmentation regime than in training, then determining the correlation of each model's training and test prediction to filter out overfitting models. Moreover, creating the ensemble by averaging the model predictions and performing thresholding on the resulting predictions and finally, Gram-OOD* adaptation, which improved OOD detection by replacing the method's generated predictions in the already created ensemble.

It is important to mention that while some of these strategies are not novel in and of itself, when combined, they have proven to be resilient for generalization in both skin lesion classification and, in particular, melanoma diagnosis.

3.1.1 Ensemble Model Selection based on Mean Correlation Matrix

The goal of our strategy inspired in [20] is to exclude models whose mean correlation of predictions revealed a significant gap between training and test predictions of the other models. The basic idea is to find the correlation between the training and test predictions for each individual model, and compute the difference in the arithmetic mean on each class correlation. Equation (1) indicates the class-wise C correlation coefficients ρ for each model which stacked form a matrix; v the validation data predictions from the training set, and t the unseen test data predictions. Equation (2) shows the Mean Correlation Matrix (MCM) which corresponds to the arithmetic mean computation of the absolute gab difference G_C of the class-wise correlations.

$$\rho(x,y)^{C}_{v|t} = \frac{\sum[(x_i - y)(y_i - y)]}{\sigma x * \sigma y} \quad (1)$$

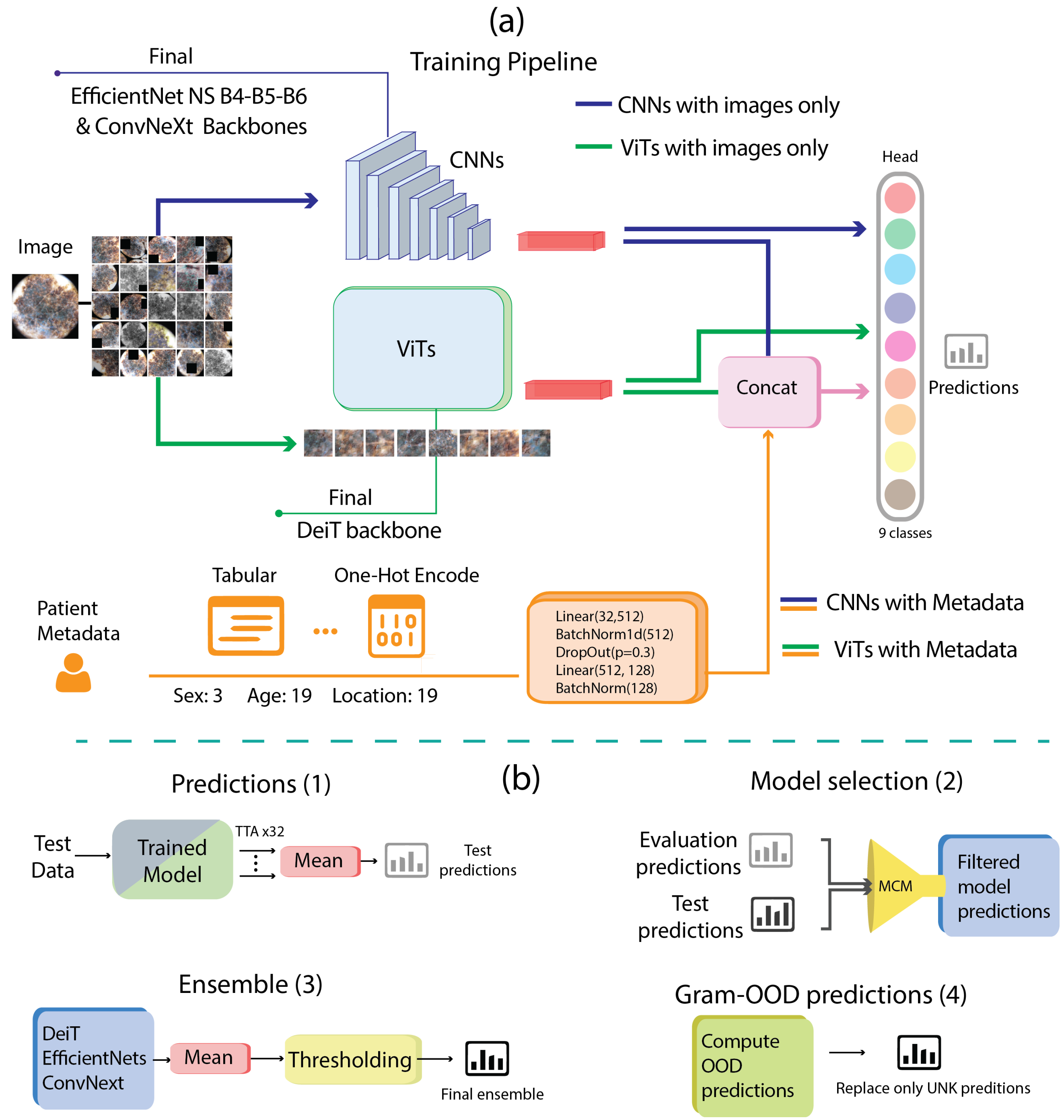


Fig. 1. Diagram of the pipeline of our model. (a) depicts the training pipeline and (b) testing pipeline. The final ensemble is trained on the datasets addressed in Table 1, and had used both external images only, and metadata for both networks. The testing pipeline shows the generation of predictions in four stages

$$MCM = \frac{1}{9}\sum_{C=1}^{9} G_C, \qquad G_C = \left|\rho_v^C - \rho_t^C\right| \tag{2}$$

As a result, a higher gap G_C indicates that model predictions behave differently between local validation and test set for that particular class. Therefore, it is possible that a feature on which this model largely depends, has a different distribution between training and test data, causing it to overfit the local data and affect generalization. Section 5.9 shows the application of the MCM in the model selection.

3.1.2 Out of Distribution with Gram-OOD

Gram-OOD, a cutting-edge Out of distribution approach that does not require extra data, was chosen to treat the OOD samples. However, as an adaptation of the Gram-OOD*, it was discovered that computing feature maps from convolutional layers rather than activation functions (see Table 2), could result in a slight improvement while retaining the pairwise correlations and layerwise deviation computation from the original method [21] and the normalization extension proposed in [22].

Method	**TNR**	**AUROC**	**DTACC**	**AUIN**	**AUOUT**
Gram-OOD* [22]	7.028	45.456	51.311	18.628	78.163
Gram-OOD (Ours)	**9.226**	**59.414**	**57.083**	**26.793**	**83.205**

Table 2. Comparison of the usage of convolutional layers vs the activation functions as feature maps

3.1.3 Imbalanced Data

As in many medical image datasets, data imbalance is a common, yet challenging issue to be addressed for model training and hyper-parameters optimization. The most popular approaches, such as Weighted Cross Entropy (WCE) [23], and Focal loss (FL) [24], were addressed in order to find the best pipeline, see Table 10. Equation 3 depicts the BCE loss function l; x represents the input, y the target, w is the weighting factor, C the number of classes and N the minibatch.

$$l(x,y) = L = l_1,...,l_N{}^\top, \qquad l_n = -\sum_{c=1}^{C} w_c log \frac{exp(x_{nc})}{\sum_{c=1}^{C} exp(x_{n,i})} y_{n,c} \tag{3}$$

Equation 4, shows the FL, were γ is the parameter for tuning, $(i-p_i)^\gamma$ the modular factor introduced to the Cross Entropy (CE), and α_i represents the weighting factor defined in practice for the FL.

$$FL = -\sum_{i=1}^{n} \alpha_i (i-p_i)^\gamma log_b(p_i) \tag{4}$$

However, the approach that consistently reached the best scores was achieved by re-scaling the output class probabilities with method known as rescaling or thresholding [25]. This approach applied in [15] has demonstrated to significantly increase the performance in imbalanced datasets by a class probability distribution approximation. [26] has showed that NNs classifiers derive Bayesian a posteriori probabilities; where they are computed for each class by their frequency in the imbalanced dataset. In other words, the output for class c for a given datapoint x implicitly corresponds a conditional probability in equation (5), where $|c|$ is the number of unique instances in class i and $p(x)$ is considered constant assuming all data have the same probability to be selected:

$$p(c|x) = \frac{p(c)p(c|x)}{p(x)}, \qquad p(c) = \frac{|c|}{\sum_k |k|} \tag{5}$$

Thus, depending on the datasets that are considered, the re-scaling made by the class probability distribution will change. Nonetheless, in order to have consistent results, the large amount of data provided by the 2019 and 2020 datasets gave a fixed set of probabilities for each class, which were used for the re-scaling factor.

4 Model Implementation

In this section, all the model implementation details and choices made are addressed.

4.0.1 Optimizer and Learning Rate

After the comparison of state-of-the-art optimizers in [27], AdamP had shown to outperform the vast majority of Gradient Descent Based optimizers in both computational cost and performance on ImageNet. Additionally, AdamP has shown advantage in a low training time context. The Learning Rate (LR) is often chosen after empirical processes and is determined by a variety of factors such as the data, models, schedulers and the optimizer itself. Nonetheless, when it comes to selecting Adam's hyper-parameters, the ML community has done a lot of experimentation and by far $3e-4$ had resonated strongly [28].

However, before making a choice on the LR, the configuration had to be selected based on the LR scheduler from section 4.0.2. In [29] the authors suggested testing any of the $3e-4$, $1e-4$, $3e-5$ as the maximum LR, and in order to have uniformity for all test, $3e-4$ was selected as the max LR.

4.0.2 LR Scheduler

Super-convergence was present in parallel throughout the whole model implementation, with the reason being it was strictly necessary given the GPU and training time limitations. However, a comparison had to be made in order to find the best super-convergence technique that could fit the project needs. The existence of super-convergence is relevant to understanding why deep networks generalize well. The ”One-Cycle” learning rate policy described in [29] requires defining a minimum and maximum LR, to achieve the super-convergence. One cycle is shown in Figure 2, that consists of two-step sizes: one in which LR increases from the min to max and the other in which it decreases from max to min of the overall number of epochs. Although other optimizers and schedulers were tested, AdamP with the One-Cycle scheduler gave the best results in the experiments. Appendix 5.4 shows the other LR schedulers tested.

4.0.3 Data Preparation

The images in the dataset are all from different sources, scanned at various resolutions and on the same color space.

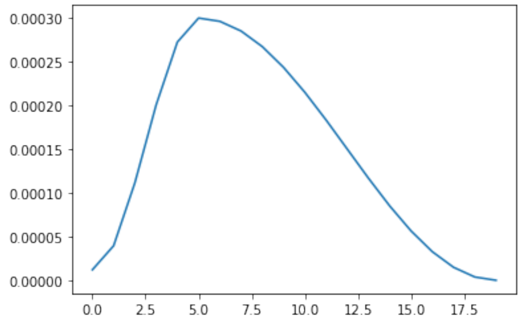
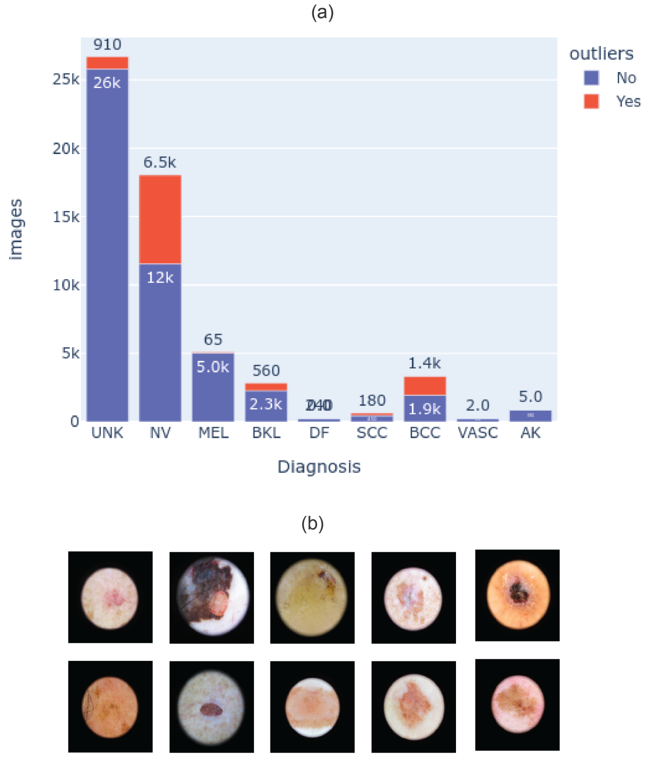


Fig. 2. One Cycle LR.



(a)

(b)

Fig. 3. (a) Outliers distribution per class; (b) a few examples of the outliers found.

However, some of them are composed of microscope-like image cropping that were detected as outliers in Figure 3, using the mean and standard deviation from the intensity values, and were preprocessed to see whether they could result in an generalized improvement as [30] stated. The data handling first consisted of trimming and cropping these microscope-lesion images, which were typically high resolution. This process resulted in another image with a lower resolution than the original, but with the item of interest (skin lesion) clearly visible and in greater detail. Figure 4 presents a few examples of all the 9577 images determined as outliers.

Additionally, it was essential to remove the missing values that were discovered during the metadata preparation. These missing values were handled by utilizing a new parameter *unknown* for the sex, age, and anatomical location. Since no newborns were recognized as patients in this dataset, the value unknown for the age was replaced as zero. Nonetheless, using the average as a mapping parameter may be an option to consider. Table 3 depicts the amount of parameters for the tabular data.

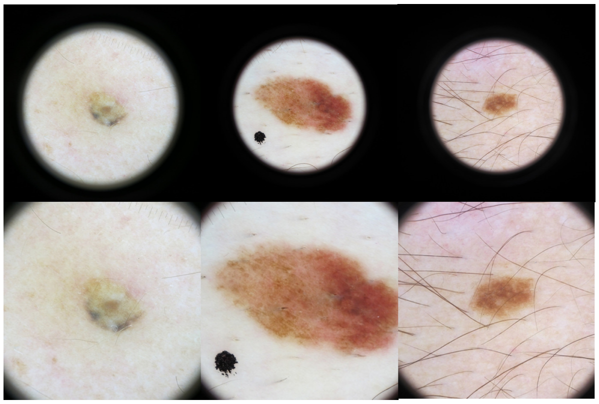


Fig. 4. Preprocesssing of outlier images.

Metadata	Number of parameters
Sex	3
Age	19
Anatomical Location	10

Table 3. Metadata number of parameters used as input for the models

4.0.4 Data Augmentation

The goal of using image augmentation is to enhance the variety in the training data with the purpose of strengthening the model's capacity to generalize. In an ideal world, there is enough training data to represent every potential variation. Nevertheless, in practice, the amount of data is a constant limitation that must be overcome. Three popular methodologies from the literature were evaluated in order to discover a suitable data augmentation regime for such real-world classification task; namely, AutoAugment [31], RandAugment [32] and AugMix [33]. Before a selection, an adaptation of the customized standard augmentation by [34] was compared in the Table 9, to find the most suitable augmentation technique in order to carefully craft the newly generated images in order to improve performance on newly unseen data.

Figure 5 shows the augmentation regime used for all the models, which was based on the idea of avoiding the deconstruction of features and patterns in the melanocytic images described in the ABCD rule [35]: where skin lesion asymmetry is a major indicator of malignant melanoma, in contrast to benign pigmented skin lesions, which are normally round and symmetric, melanomas spread uncontrollably. As a result, asymmetry, border, color, and diameter are critical in developing a skin lesions augmentation regime. Taking

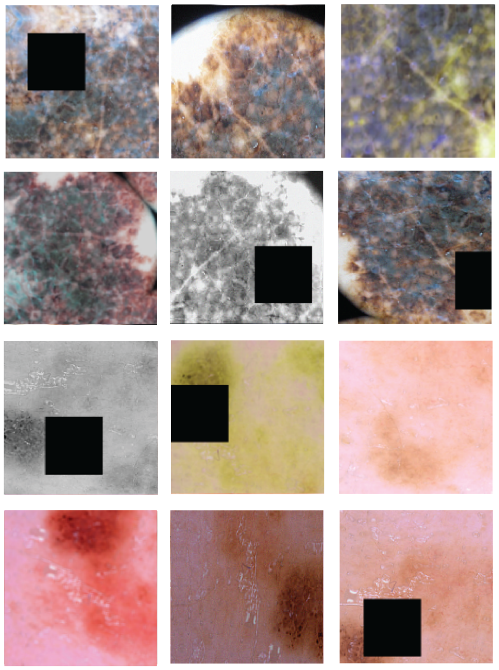


Fig. 5. Image augmentation employed: a standard augmentation regime (random flip, rotation, brightness/contrast and blur/gaussian noise) followed by a random and resized crop strategy, CutOut of 30% image size, and gray and color-jitter/hue-saturation changes. Details can be found in 8.

inspiration from Contrastive Learning [36] the composition of simple augmentations for learning good representations, gray and color distortions were adopted. Moreover, key to the locality of the augmentation was a heavy cropping strategy, where random resized crops were fed into the models followed by random brightness and contrast changes including color jitter, random flipping, random rotation, random scaling, and random blur/noise/sharpen changes. Furthermore, CutOut [37] was used with one hole that was 30% the size of the image and had a 50% chance of appearing. Finally, a couple of augmentation strategies, including microscopy-crop and color constancy shades of grey as in [30], were explored, but yielded no benefits and were therefore rejected. Table 8 has the whole augmentation configuration tested.

5 Validation

In this section, first we discuss the evaluation metrics, followed by the main framework and tools. We show the baseline and default settings, as well as the data splitting. Furthermore, we illustrate the best data augmentation and the imbalanced data methods followed by the final results on both challenges where we achieved the first place on the ISIC 2019 to date, and the third place on the ISIC 2020.

5.1 Evaluation Metrics

In order to assess the model performance and compare different models we used the following metrics: Accuracy (ACC), sensitivity (SE), specificity (SP), Dice Coefficients (DI) and Area Under the Curve (AUC) score. In Table 4, the formulas for these metrics are presented. Another common method for examining how probabilistically the model yields results is the receive operating characteristics (ROC) curve which displays the ratio of true to incorrect predictions.

Additionally, for the genralization evaluation, the automatic scoring system available for the 2019 ISIC Challenge uses the following norms [38]:

The validation score is computed with the goal metric (balanced multi-class accuracy), taken against a small (100), non-representative, pre-determined subset of images.

For reference, a random submission generates a validation score of about 0.3.

Diagnosis confidences are expressed as floating-point values in the closed interval [0.0, 1.0], where 0.5 is used as the binary classification threshold.

The image field uses values with an ISIC_ prefix and without any .jpg file extensions

The values are floating point (0 and 1 are invalid, but 0.0 and 1.0 are valid)

The row values do not necessarily sum to 1.0

The greatest value of each row is considered the overall diagnosis prediction

All values greater than 0.5 are considered positive binary diagnosis predictions

Metric	Formula
Accuracy	$ACC = \frac{TP+TN}{TP+FP+TN+FN}$
Balanced accuracy	$BACC = \sum_{1}^{N} \frac{ACC_c}{N}$
Sensitivty	$SE = \frac{TP}{TP+FN}$
Specificity	$SP = \frac{TN}{TN+FP}$
Precision	PE $P = \frac{TP}{TP+FP}$
Average precision	AP$AV = \frac{TP}{TP+FP}$
Dice coefficient	DI$\frac{2 \cdot TP}{2 \cdot TP+FN+FP}$
Area under the curve	$AUC = \int_{0}^{1} TP(FP) \delta FP$

Table 4. Metrics defined by the 2019 ISIC live challenge to assess models performance.

5.2 Framework and Tools

DL has evolved swiftly from basic feed forward layers to complex numerical algorithms. Performance is crucial in collaboration with a dynamic eager execution to facilitate work for data scientists, researchers, and students. As a result, popular tools like Pytorch have emerged [39], which include a novel ecosystem for applying DL with a focus on performance and a usability centric design, where the DL models are seen as python programs that perform immediate execution of dynamic tensor computations and GPU acceleration.

In relation to GPUs, depending on the availability of the Google Colab —our computing resource—, over 300 models were trained on a variety of GPUs, including the Tesla T4-16GB, Tesla P100-PCIE-16GB and Tesla V100-SXM2-16GB. Nonetheless, in order to get access to the powerful GPUs and longer runtime notebooks, we proceeded with a Colab Pro+ subscription for 4 months and a Colab Pro subscription for a 3 months. The lack of GPU availability, combined with the limitation of notebook runtime to a maximum of 24 hours before they are shut down, resulted in a training limitation that had to be overcome by designing a pipeline in which super-convergence was the key element to achieve competent results and train from one to two daily models. Finally, the accessibility given by with timm's library [40], of a broad range of cutting-edge ViTs and CNNs models with pre-trained weights, along with key tools and frameworks such as Pytorch Lightning [41], and Wandb [42], enabled the management of multiple experiments, running them in parallel and comparing them in real-time, greatly accelerating results evaluation and tracking, assessing which augmentation regimes and hyperparameter changes were yielding positive results, and having the best models available to run the TTA predictions whenever GPUs were available.

Our code is publicly available under MIT License at https://github.com/blobquiet/SIIM-ISIC-Melanoma-Classification. Email: blobquiet@gmail.com

5.3 Baseline and Default Settings

In order to get a decent start, the results from the CNNs baseline in research [15] were adopted and with the configurations made from the training and computational limitations mentioned in Section 5.2. Furthermore, a baseline of ViTs had to be obtained in order to have a first look and comparison between ViTs and CNNs in the skin lesion classification task. The CNNs that were used for baseline comprise the Efficient Nets [17], Inception Resnet V2 [43] and ResNeXt [44]. In the case of ViTs used as baseline: the basic ViT [9], BEiT [45], SwinT [46], and SwinTV2 [47]. Hence, the relevant models and their performance are displayed in Table 5. Furthermore, initially only images from the whole dataset shown in Figure 6 were used. As a result, 29,639 training samples and 3296 validation images were used with a 90-10 split from the PH2, 7 point criterion, MED-NODE, SKINLV2-V1-2-3, SD-198, and ISIC 2019 datasets; melanoma had 4914 samples for baseline.

With this particular setup, preliminary results show that CNNs defeat ViTs ensemble by a narrow margin. One key point to note is that the image size was multi resolution, and the EfficientNet B5 received the highest score of 0.483. Because ViTs lacked the richness of scaled resolution, a diversity of input sizes for the ViTs backbones is required to assess the outcomes properly.

A key finding from Table 5, was that the Swin transformer outperformed all of the CNNs excluding the EfficientNet B5. This might be attributed to the locality of CNNs when processing the raw dataset, as in some image crops, the network may be fed a fully or almost entirely black image from the microscope circular mask, as found in Figure 3, and a large image size can counter that by assuring that there will always be information in the random crops which explains the high score from the EfficientNet B5.

5.3.1 Default Settings

Following isolated experiments, all models were trained using fine-tuning on 10 epochs in 16-bit mixed-precision, with a batch size of 32, and using gradient accumulation when necessary. The optimizer and LR scheduler that performed best given the computing constraints were the One Cycle LR discussed in Section 4.0.2, and AdamP 4.0.1, with the recommended learning rate of $3e-4$. Additionally, both training and evaluation were carried out matching the same model input resolution. Finally, before averaging the predictions, TTA was applied 8, 20 and 32 times without CutOut and the data augmentation regime was chosen following the methodology described in Section 4.0.4. The final configuration of our ensemble is given in Table 6.

5.4 Super-convergence, Optimizers and Schedulers

In this section, the primary goal was to assess the phenomenon of Super-Convergence using a variety of popular optimizers and schedulers. Experiments were made to tune and find the most suitable optimizer and LR scheduler. The first consisted in comparing the OneCycle LR with four optimizers; Cosine Annealing and SGD Cosine Annealing with warm restart [49], SGD with Cyclic LR [50], and straightforward LR step decay with AdamP. The strategies assesst in the experiments can be found in Table 7. The trials have revealed that the OneCycle LR is the best candidate for further testing given that it produces by far the best results of all in a 10-epoch training session. It should be noted that their assessment was conducted using the identical LR $3e-4$ and a middle ground image size of 380 from the EfficientNet-B4 backbone.

5.5 Data Splitting

For the data splitting the objective was find a strategy that could work for both model selection and hyperparameter optimization. The holdout method is the simplest strategy for evaluating a classifier and although it is not the best strategy to exhaustably assess the models on the whole bulk of the data, it provides the advantage of immediate experiments to determine the fundamental settings for a robust classifier. To achieve generalization on previously unseen data, it was vital to verify that the training and validation were representative of the full dataset. As a consequence, a stratified split based on the skin lesion target class was necessary, and based on the empirical findings, a 90% to 10% split was decided. Following a data-driven approach, adding external data as in [15], demonstrated a slight improvement for the outlier class. Therefore, datasets described in Figure 6 were used to feed the models in order to reach diversity

Method	# Params	Image size	Data usage	Val BACC	2019 Score
SWSL ResNeXt-101 32x4d [48]	54M	224	External	72.09%	0.429
Inception-ResNet-V2 [43]	56M	299	External	76.33%	0.433
EfficientNet b4 [17]	19M	380	External	71.11%	0.424
EfficientNet b5 [17]	30M	456	External	77.73%	**0.483**
CNNs baseline ensemble	**0.496**				
ViT-L-16 [9]	304M	224	External	75.73%	0.418
Swin-L-4 [46]	197M	224	External	73.02%	**0.464**
SwinV2-B- [47]	88M	256	External	74.56%	0.412
BeiT-B-16 [45]	87M	224	External	75.13%	0.403
ViTs baseline ensemble	0.482				

Table 5. ISIC 2019 score and BACC baseline. Note that there is no data preprocessing, duplicates removal or imbalance handling.

Settings	Model
Pretrained	True
Fine-tunning	unfreezed layers from start
Image Size	Same as backbone
Optimizer	AdamP
Weight decay	0
Momentum	$B1$,$B2$=(0.9,0.999)
Batch size	32 with Gradient accumulation when needed
Learning rate Scheduler	OneCycle LR
Anneal strategy	Cosine
Base momentum	0.85
Max momentum	0.95
Max LR	3e-4
Max epochs	10
Mixed precision	16 bits
TTA	32
Augmentation	[5.6]

Table 6. Training and hyper-parameter configuration for the final models

Scheduler	Model	Optimizer	Epochs	Val BACC
Cosine Annealing	EfficientNet B4	AdamP	20	88.44%
Cosine Annealing warm restart	EfficientNet B4	SGD	20	86.25%
Cyclic LR	EfficientNet B4	SGD	20	82.77%
Step LR	EfficientNet B4	AdamP	20	76.34%
OneCycle LR	EfficientNet B4	Swin-L-4	20	**90.19**%

Table 7. Optimizer and LR scheduler experiments in order to find the best approach for super-convergence.

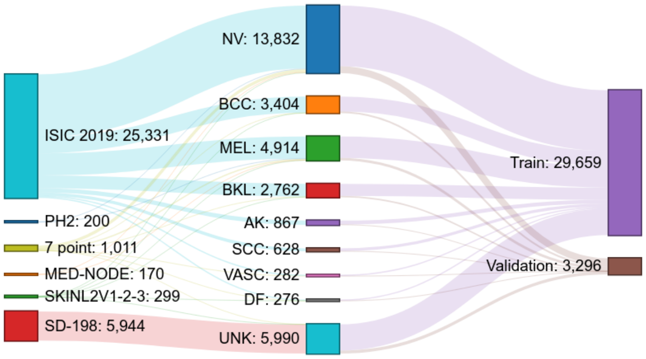


Fig. 6. Skin Lesion Datasets Distribution for the external data. It displays the 25,331 samples from the ISIC 2019 as well as the contributions from the remaining external datasets and also indicates the splitting made for training and validation.

in the DL ensemble. Moreover, in order to include metadata features from section, the ISIC 2019 and ISIC 2020 datasets were both used for training with bulk of 57301 images. The stratified split can be inspected in Figure 7.

As a side note, if the goal is to improve the general-

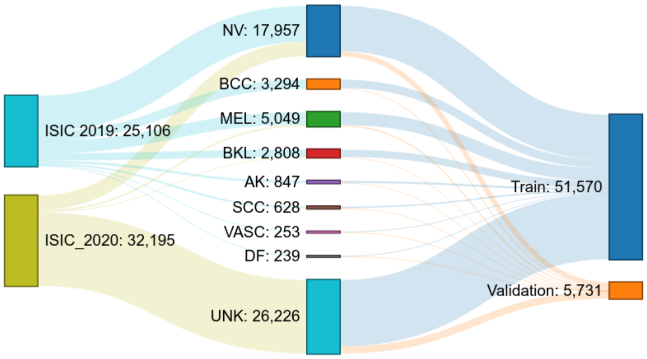


Fig. 7. Metadata Skin Lesion Datasets Distribution for the 2019 and 2020 ISIC datasets. The contribution in each class is clearly demonstrated here, along with the splitting approach and proportions for training and validation.

ization performance and time or computing resources are not a constraint, stratified K-fold Cross Validation (CV) is a suitable data splitting strategy for model selection [51]. Nonetheless, the fact that just by using the hold-out method our model could achieve top scores, demonstrates the potential and possibility of increasing the score further on the live leaderboard by simply integrating CV for model selection.

5.6 Data Augmentation Validation

We applied 13 techniques of data augmentation that are shown in Table 8. Table 9 compares the augmentation regime employed in this study to the alternative conventional augmentation methods. The experiments used the same setup as in the baseline 5.3, with both the ViT and the CNN, and the same image size of 224. It is important to note that no thresholding, WCE, or FL were applied in this experiment to provide a raw perspective of the results, which explains lower results. Therefore, only regular CE was employed in this experiment, which served to determine which augmentation strategy works best.

Following the criteria from the melanoma ABCD rule [52], the Adapted Augmentation regime produced the best overall results, with higher validation Balanced Multiclass Accuracy of 83.62% and 83.87% and an overall score of 0.495 and 0.479 for the EfficientNetV2-B0 and Swin-L-4, respectively. As a result, the data augmentation regime was employed for all following research.

5.7 Imbalanced data method comparison

The purpose of this experiment was to show that our rescaling method to treat better the imbalanced data compared to weighted cross entropy and focal loss. We show the experiments using two of the baseline models in particular, a CNN and a transformer. Table 10 presents the comparative results of the three techniques indicated in 3.1.3 in order to analyze which approach among the conventional methods for handling imbalanced datasets in skin lesion classification should be preferred.

The tests were carried using the two networks from the preceding section, both CNNs and ViTs. These show that thresholding beats the other two by a significant margin, ranging from 0.022 with the WCE to 0.011 with FL. As a result, the thresholding strategy was adopted after the predictions, implying that the non-weighted CE had to be used as a loss function for training, and thresholding applied at inference from this point on.

5.8 ViTs and CNNs Ensemble Results for 2019 ISIC Challenge

The 2019 ISIC Challenge, which contains an automatic scoring system and 8,239 challenging images in the test set, allowed for credibility in the evaluation of our model's generalization capabilities. The top network results, which were obtained through an ensemble of the ViTs and CNNs, are shown in Table 11. Although BEiT-L is a powerful network for the ImageNet dataset, as demostrated by [45], it underperformed in all of the test results from ViTs —with less than 0.500 for ISIC 2019 test score after thresholding— and hence had was omitted. Additionally, Table 12, depicts the ensemble methods chosen by verifying with three possibilities: (1) a rank of probabilities, (2) majority voting scheme, and (3) model averaging; with the averaging yielding 0.600 and overall the best results from the comparison.

Furthermore, the ensemble predictions were created using only the top six models from ViTs and CNNs. Although the 384 image size was best for the ViTs and the 380 image resolution was best for the CNNs, the multi-resolution technique for ensemble diversification allowed us to construct ensembles that outperformed any of the individual models ranging from 224 to 528. The DeiT-D3 in particular achieved a top validation score of 91.73% and a high score of 0.593, indicating that it had capture features not present in the other models. CNNs, on the other hand, outperform ViTs for the majority of individual ensembles in both external and meta data. Finally, it was not intended to utilize a brute force averaging strategy, as was the case in earlier 2019 and 2020 ISIC submissions, hence a model selection approach had to be used.

In order to take explicit care of OOD samples and outperform the current methods in the challenges, we used the Gram-OOD to calculate the OOD samples, as shown in Section [22] and described in Section 3.1.2.

Table 13 depicts a comparison after the Gram-OOD method was applied, accounting for a slight improvement in the AUC. We achieved AUC sensitivity higher than 80% and average precision with 0.686, 0.437 and 0.302, respectively.

Finally, the outlier class improvement is shown in Figure 8. It illustrates the new ROC Curve for the UNK class, alongside a dashed line corresponding to the previous ROC Curve (a) from Figure 12. The rest of the classes remain the same as the Gram-OOD only replace the predictions from the outlier unknown class.

5.9 Model Selection

Once the previous results have achieved second place in the ISIC 2019 live leaderboard with the CNN ensemble, the

Augmentation functions **(From Albumentations)**	**Detalis**

RandomResizedCrop	height = size, width = im_size
	scale=(0.08,1.0), ratio=(0.75,1.3333)
	interpolation = cv2.INTER_CUBIC, p = 1

Rotate	p=0.5
Flip	p =0.5
Affine	mode=4, p=0.5

ColorJitter	brightness=0, contrast=0
	saturation=0.3, hue=0.1, p = 0.5

Transpose	p=0.5
ToGray	p=0.2

RandomBrightnessContrast	brightness_limit=0.2,
	contrast_limit=0.2, p=0.5

HueSaturationValue	hue_shift_limit=2, sat_shift_limit=15
	val_shift_limit=20,p = 0.5

ShiftScaleRotate	shift_limit=0, scale_limit=(0.0, 0.05)
	rotate_limit=0, interpolation=1,
	border_mode=0, p=0.5

One of
Blur, GaussNoise, IAASharpen
	Blur(blur_limit=5, p=0.3),
	GaussNoise(var_limit=(5.0, 10.0), p=0.3)
	IAASharpen(alpha=(0.1, 0.3), lightness=(0.5, 1.0), p=0.4)

Cutout	max_h_size=int(im_size*0.375),
	max_w_size=int(im_size*0.375), num_holes=1, p=0.5

Normalization	mean=(0.485, 0.456, 0.406)
	std=(0.229, 0.224, 0.225

Table 8. Albumentation configuration for the training data.

Method	**Model**	**Metric**	
		Val BACC	**2019 Score**
AugMix [33]	EfficientNetV2-B0	78.83%	0.429
	Swin-L-4	76.14%	0.403
AutoAugment [31]	EfficientNetV2-B0	79.35%	0.434
	Swin-L-4	77.67%	0.552
RandAugment [32]	EfficientNetV2-B0	80.07%	0.439
	Swin-L-4	79.6%	0.419
Adapted Augmentation (Ours)	EfficientNetV2-B0	83.62%	**0.495**
	Swin-L-4	83.87%	**0.479**

Table 9. Data augmentation comparison results, using a ViT and a CNN for each data augmentation regime.

best models to enhance the ensemble for ViT must be identified. The approach for determining the optimal ensemble is provided here, which entails assessing a gap between models using the correlation of training with test predictions for each model. Therefore, MCM as used by [20], was extended in this study for the nine class predictions (see Section 3.1.1).

Figure 9 illustrates the results gap generated to select the models selected for the ensemble. It is worth noting that the Deit-L appears to be among the most feature-rich model, with an overall gap of 0.42, followed by the ConvNext-B with a 0.45. As a result, these two models were chosen for the ensemble; it is noticeable that the EfficientNets with Noisy Student weights outperformed the ViTs in the task as a backbone; the B4, B5 and B6 gaps are the ones that follow

Imbalanced method	Model	Metric	
		Val BACC	2019 Score
Weighted Cross Entropy	EfficientNetV2-B0	84.34%	0.511
[23]	Swin-L-4	81.94%	0.504
Focal Loss	EfficientNetV2-B0	87.24%	0.521
[24]	Swin-L-4	86.75%	0.515
Thresholding	EfficientNetV2-B0	86.94%	**0.536**
(Ours)	Swin-L-4	86.75%	**0.526**

Table 10. Comparison of experimental results for Imbalanced methods

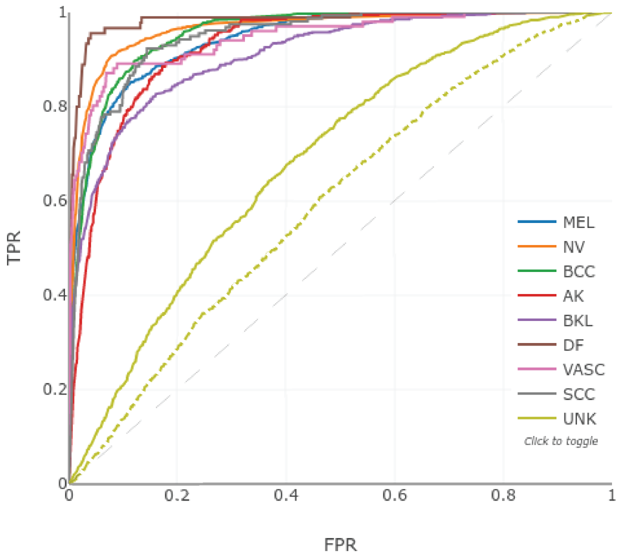


Fig. 8. ROC curve with improvement AUC for the unknown class

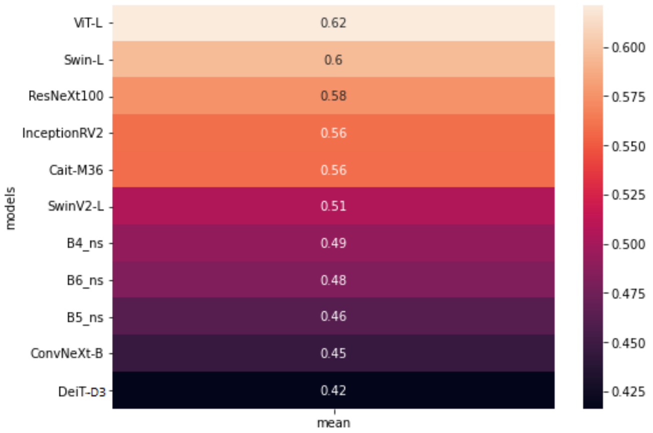
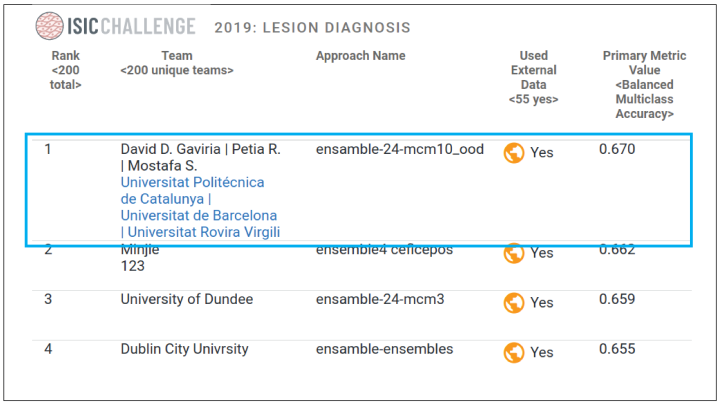


Fig. 9. Mean correlation matrix of predictions for model selection. The higher gap means a poor model, likely overfitting on local data.

with 0.46, 0.48 and 0.49 respectively. Finally, the remaining models were eliminated one by one since it was determined that each one was degrading the total score.

5.10 ViTs and CNNs Final Ensemble

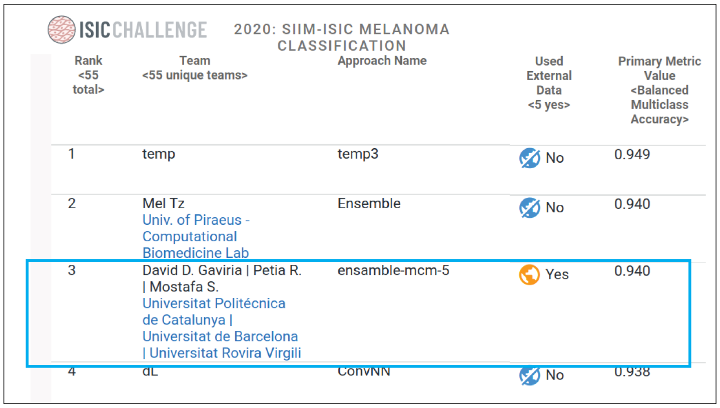
Table 14 represents the ensemble that had reached first place in 2019 ISIC live challenge and third place in 2020 ISIC live challenge, see Figures 10 and 11. It was composed of a diversification of models, both ViTs and CNNs in Table 11, and discriminated after a model selection with the MCM from the previous section 5.9.



2019: LESION DIAGNOSIS

Rank <200 total>	Team <200 unique teams>	Approach Name	Used External Data <55 yes>	Primary Metric Value <Balanced Multiclass Accuracy>
1	David D. Gaviria | Petia R. | Mostafa S. Universitat Politécnica de Catalunya | Universitat de Barcelona | Universitat Rovira Virgili	ensamble-24-mcm10_ood	Yes	0.670
2	Minjie 123	ensemble4 cefícepos	Yes	0.662
3	University of Dundee	ensamble-24-mcm3	Yes	0.659
4	Dublin City Univrsity	ensamble-ensembles	Yes	0.655

Fig. 10. First place in 2019 ISIC live leaderboard



2020: SIIM-ISIC MELANOMA CLASSIFICATION

Rank <55 total>	Team <55 unique teams>	Approach Name	Used External Data <5 yes>	Primary Metric Value <Balanced Multiclass Accuracy>
1	temp	temp3	No	0.949
2	Mel Tz Univ. of Piraeus - Computational Biomedicine Lab	Ensemble	No	0.940
3	David D. Gaviria | Petia R. | Mostafa S. Universitat Politécnica de Catalunya | Universitat de Barcelona | Universitat Rovira Virgili	ensamble-mcm-5	Yes	0.940
4	dL	ConvNN	No	0.938

Fig. 11. First place in 2020 ISIC live leaderboard

5.10.1 ISIC Submissions and Evaluation

We submitted our model to the ISIC Challenge submission system, which allows for automatic format validation and scoring explained in 5.1. Figure 12 and Table 14 resume the results obtained from the unseen data for the 2019 Challenge and the 2020 ISIC challenge: (a) shows the ROC Curve result for each individual class in the 2019 challenge, and (b) shows the melanoma predictions results illustrated in the ROC Curve from the ISIC 2020 dataset.

A brief look at Figure 12 ROC curve and AUC reveals that the ROC curve performs much worse with the UNK class than with the other classes. Likewise from Table 14,

Method	# Params	Image size	Data usage	Val BACC	2019 Score
ViT-L-16	26M	224	External	78.35%	0.514
[9]			Meta	83.56%	0.527
VOLO-D3	306M	512	External	82.31%	0.512
[53]			Meta	85.36%	0.516
DeiT-D3	305M	384	External	89.97%	0.592
[16]			Meta	**91.73%**	**0.593**
CaiT-M-36	271M	380	External	84.29%	0.571
[54]			Meta	88.21%	0.589
Swin-L-4	197M	224	External	81.17%	0.526
[46]			Meta	83.87%	0.564
Swin-L-V2	197M	384	External	86.10%	0.563
[47]			Meta	89.46%	**0.610**
ViTs Ensemble	0.612				
SWSL ResNeXt-101 32x4d	54M	224	External	75.73%	0.576
[48]			Meta	74.06%	0.579
Inception-ResNet-V2	56M	299	External	78.23%	0.586
[43]			Meta	78.25%	0.587
EfficientNet b4 NS	19M	380	External	83.66%	0.603
[18]			Meta	84.85%	**0.630**
EfficientNet b5 NS	30M	456	External	78.25%	0,604
[18]			Meta	85.94%	0.618
EfficientNet b6 NS	43M	528	External	85.99%	0.612
[18]			Meta	86.07%	**0.630**
ConvNeXt-B	89M	384	External	85.91%	0.592
[19]			Meta	86.95%	0.594
CNNs Ensemble	**0.660**				

Table 11. Balanced Multiclass Accuracy of training in ViTs and CNNs state-of-the-art models. All hold-out splitting with 90 to 10% for training and validation. It was considered a heavy cropping strategy with TTA 32 and only 10 epochs training via fine-tuning. Values are given in percentage as validation of the BACC. Ensemble was used as the average of all predictions from ViT and CNN models. External refers to both the 2019 dataset and the external datasets, and Meta means the 2019 dataset and 2020 datasets training both the images and metadata. In all cases, the nine classes were used for prediction

Ensemble method	ViTs ensemble	CNNs ensemble
	2019 Score	2019 Score
Rank of probabilities	0.611	0.647
Majority voting	0.542	0.603
Averaging	**0.612**	**0.660**

Table 12. Ensemble method used for both the ViTs and CNNs

all classes have an AUC greater than 0.9, with the exception of the outlier class, which has the lowest AUC of 0.595. Nonetheless, specificity with a score of 1 for the UNK means that the model has correctly identifying all the negative predictions for the outlier class, but in contrast, the true positive rate calculated by the sensitivity had a score of zero, indicating that the outlier class was unable to classify any of the positive samples. Overall, the results account for the challenging task of classifying OOD samples.

Moreover, in the case of melanoma, the AUC from table 14 shows a competent score of 0.943 which motivated a submission in the 2020 ISIC challenge that assesses the malignant prediction.

The ROC Curve (b) in Figure 12 and the metrics results in Table 15 are the results of the submission to the 2020 ISIC live challenge. The 0.940 AUC allowed the project to finish third in the 2020 ISIC live challenge, confirming the proposal's generalization capabilities in a different test dataset.

6 Conclusions

The study proves that despite the fact that not a single model, nor ViTs or CNNs could achieve a very high standing in both the 2019 and 2020 ISIC live challenges, an ensemble of ViTs and CNNs was able to provide a huge diver-

Metric	AUC	AUC Sens >80%	Average Precision	Accuracy	Sensitivity	Specificity	Dice Coefficient	PPV	NPV
Unk	0.595	0.310	0.234	0.808	0.00	1.00	0.00	1.00	0.808
Unk-OOD	**0.686**	**0.437**	**0.302**	0.808	0.00142	1.00	0.00283	1.00	0.808

Table 13. Outlier class metrics comparison with the OOD results for the top 1 in the 2019 ISIC live challenge

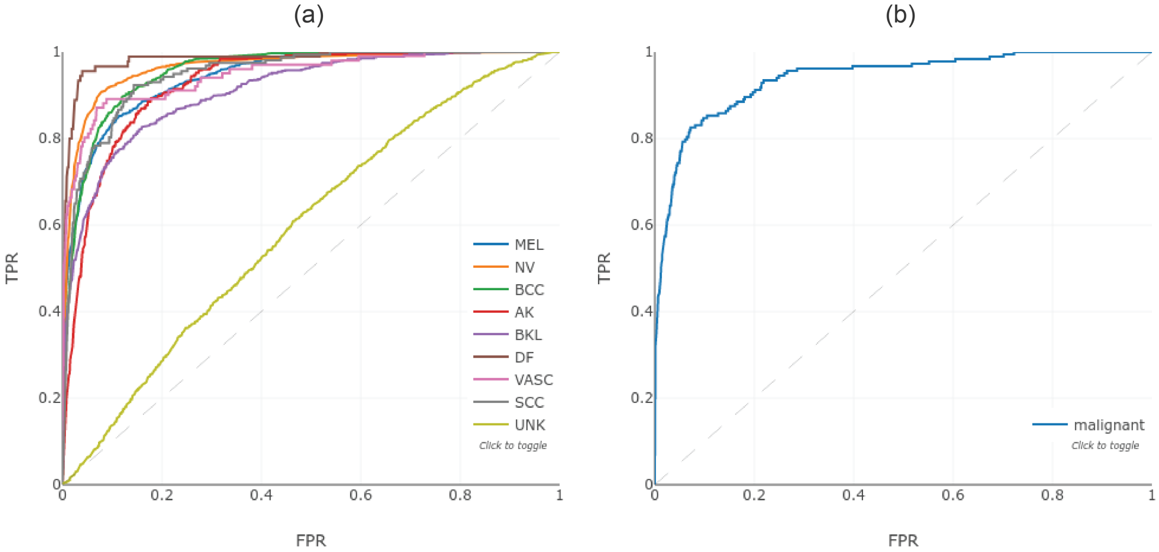


Fig. 12. ROC curve for (a) the 0.670 balanced multi-class accuracy ensemble for the 2019 ISIC Challenge and (b) the melanoma with 0.940 AUC for the 2020 ISIC Challenge.

Metrics	Mean	Diagnosis Category								
		MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
AUC	**0.908**	**0.943**	0.965	0.955	0.928	0.911	0.983	0.947	0.949	0.595
AUC, Sens >80%	**0.836**	**0.892**	0.943	0.915	0.861	0.820	0.975	0.918	0.887	0.310
Average Precision	**0.597**	**0.821**	0.938	0.774	0.404	0.640	0.608	0.572	0.382	0.234
Accuracy	**0.928**	**0.913**	0.910	0.918	0.931	0.937	0.986	0.981	0.972	0.808
Sensitivity	**0.589**	**0.658**	0.797	0.788	0.610	0.490	0.733	0.653	0.573	0.00
Specificity	**0.972**	**0.965**	0.964	0.938	0.948	0.979	0.989	0.985	0.981	1.00
Dice Coefficient	**0.538**	**0.719**	0.851	0.716	0.474	0.572	0.559	0.482	0.471	0.00
PPV	**0.630**	**0.791**	0.913	0.655	0.388	0.688	0.452	0.382	0.400	1.00
NPV	**0.948**	**0.933**	0.908	0.967	0.978	0.953	0.997	0.995	0.991	0.808

Table 14. Ensemble metrics for top-1 in the 2019 ISIC live challenge.

Metric	AUC	AUC Sens >80%	Average Precision	Accuracy	Sensitivity	Specificity	Dice Coefficient	PPV	NPV
MEL	**0.940**	0.899	0.544	0.982	0.284	0.999	0.426	0.852	**0.983**

Table 15. Ensemble melanoma metrics for top-3 in the 2020 ISIC live challenge.

sity, necessary to achieve top-1 for the 2019 challenge and top-3 for the 2020 ISIC challenge. Although, improvements were made in the topic of outliers, both for the data-driven approach and from the Gram-OOD* adaptation, the OOD samples present in the 2019 ISIC remain an open challenge and further research on the topic is required to improve OOD detection for both CNNs and ViTs.

Based on the classification of skin lesions and recent publications, two open live challenges—ISIC 2019 and ISIC 2020— were organised to boost research in analysis of dermatological images and validate the overall performance of the deep learning solutions. Our thorough analysis of the problem led to the following observations:

(1) Using CNNs and ViTs architectures in an ensem-

ble to classify skin lesions can significantly improve disease diagnosis performance by offering a wider range of diverse predictions, reaching top-1 in the ISIC 2019 challenge;

(2) After training on all nine skin diseases, employing a particular class prediction for melanoma results in a robust generalization classifier that performs well on unseen test data;

(3) The diversity provided by the image-level and patient-level metadata is one of the responsibles for the results' improvement;

(4) Applying Bayes' theorem to predictions in an extremely unbalanced dataset can additionally enhance the model generalization.

Key findings: When classifying skin lesions, especially in melanoma appearance, is important to consider both the augmentation distortions and the patient's context; two comparable skin moles can improve feature extraction in a classifier if they are considered at the patient-level belonging to the same patient with one of them known to be malignant, but the other benign. However, an augmentation scheme that alters a skin mole to resemble a melanoma, especially when combined with elastic asymmetric transformations or a grid distortions, may seriously hinder the deep NNs learning capabilities.

Furthermore, dermoscopy is usually used for melanomas and other kinds of skin cancers with pigmentation, however, it is difficult to access a dermoscope in resource-poor regions, and it is unnecessary for most of the common skin diseases. Therefore, developing an effective skin disease diagnosis system based on easily accessed clinical images would be beneficial and could provide low-cost, universal access to more people [55]. Although, some of the data used here mixed dermoscopy and clinical images, further research is required to assess the behavior of a DL solution with a bulk of clinical images in the test set.

References

[1] WHO(2017), W. H. O. Radiation: Ultraviolet (uv) radiation and skin cancer. `https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer`. Accessed: 2022-06-30.

[2] Forsea, A. M., 2020. "Melanoma epidemiology and early detection in europe: Diversity and disparities.". *Dermatology practical & conceptual,* **10 3**, p. e2020033.

[3] Arnold, M., Singh, D., Laversanne, M., Vignat, J., Vaccarella, S., Meheus, F., Cust, A. E., de Vries, E., Whiteman, D. C., and Bray, F., 2022. "Global burden of cutaneous melanoma in 2020 and projections to 2040.". *JAMA dermatology*.

[4] Leonardi, G. C., Falzone, L., Salemi, R., Zanghì, A., Spandidos, D. A., McCubrey, J. A., Candido, S., and Libra, M., 2018. "Cutaneous melanoma: From pathogenesis to therapy (review)". *International Journal of Oncology,* **52**, pp. 1071 – 1080.

[5] ACS (2022), A. C. S. Survival rates for melanoma skin cancer. `https://www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stag` `html`. Accessed: 2022-07-30.

[6] Chang, W.-Y., Huang, A., Yang, C.-Y., Lee, C.-H., Chen, Y.-C., Wu, T.-Y., and Chen, G.-S., 2013. "Computer-aided diagnosis of skin lesions using conventional digital photography: A reliability and feasibility study". *PLoS ONE,* **8**.

[7] Belilovsky, E., Eickenberg, M., and Oyallon, E., 2019. "Greedy layerwise learning can scale to imagenet". *ArXiv,* **abs/1812.11446**.

[8] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L., 2015. "Imagenet large scale visual recognition challenge". *International Journal of Computer Vision,* **115**, pp. 211–252.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., 2021. "An image is worth 16x16 words: Transformers for image recognition at scale". *ArXiv,* **abs/2010.11929**.

[10] Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., 2017. "Attention is all you need". In NIPS.

[11] Chen, J., Chen, J., Zhou, Z., Li, B., Yuille, A. L., and Lu, Y., 2021. "Mt-transunet: Mediating multi-task tokens in transformers for skin lesion segmentation and classification". *ArXiv,* **abs/2112.01767**.

[12] Combalia, M., Codella, N. C. F., Rotemberg, V. M., Helba, B., Vilaplana, V., Reiter, O., Halpern, A. C., Puig, S., and Malvehy, J., 2019. "Bcn20000: Dermoscopic lesions in the wild". *ArXiv,* **abs/1908.02288**.

[13] Tschandl, P., Rosendahl, C., and Kittler, H., 2018. "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". *Scientific Data,* **5**.

[14] Rotemberg, V. M., Kurtansky, N. R., Betz-Stablein, B., Caffery, L. J., Chousakos, E., Codella, N. C. F., Combalia, M., Dusza, S. W., Guitera, P., Gutman, D., Halpern, A. C., Kittler, H., Köse, K., Langer, S. G., Liopryis, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A. J., Tschandl, P., Weber, J., and Soyer, H. P., 2021. "A patient-centric dataset of images and metadata for identifying melanomas using clinical context". *Scientific Data,* **8**.

[15] Steppan, J., and Hanke, S., 2021. "Analysis of skin lesion images with deep learning". *ArXiv,* **abs/2101.03814**.

[16] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and J'egou, H., 2021. "Training data-efficient image transformers & distillation through attention". In ICML.

[17] Tan, M., and Le, Q. V., 2019. "Efficientnet: Rethink-

ing model scaling for convolutional neural networks". *ArXiv,* **abs/1905.11946**.

[18] Xie, Q., Hovy, E. H., Luong, M.-T., and Le, Q. V., 2020. "Self-training with noisy student improves imagenet classification". *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695.

[19] Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., and Xie, S., 2022. "A convnet for the 2020s".

[20] Nikita Kozodoi, Gilberto Titericz, S. H. G., 2020. 11th place solution writeup. `https://www.kaggle.com/competitions/siim-isic-melanoma-classification/discussion/175624`. Accessed: 2022-04-30.

[21] Sastry, C. S., and Oore, S., 2019. "Detecting out-of-distribution examples with in-distribution examples and gram matrices". *ArXiv,* **abs/1912.12510**.

[22] Pacheco, A. G. C., Sastry, C. S., Trappenberg, T. P., Oore, S., and Krohling, R. A., 2020. "On out-of-distribution detection algorithms with deep neural skin cancer classifiers". *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3152–3161.

[23] Aurelio, Y. S., de Almeida, G. M., de Castro, C. L., and de Pádua Braga, A., 2019. "Learning from imbalanced data sets with weighted cross-entropy function". *Neural Processing Letters,* **50**, pp. 1937–1949.

[24] Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., and Dollár, P., 2020. "Focal loss for dense object detection". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **42**, pp. 318–327.

[25] Buda, M., Maki, A., and Mazurowski, M. A., 2018. "A systematic study of the class imbalance problem in convolutional neural networks". *Neural networks : the official journal of the International Neural Network Society,* **106**, pp. 249–259.

[26] Richard, M. D., and Lippmann, R., 1991. "Neural network classifiers estimate bayesian a posteriori probabilities". *Neural Computation,* **3**, pp. 461–483.

[27] Heo, B., Chun, S., Oh, S. J., Han, D., Yun, S., Kim, G., Uh, Y., and Ha, J.-W., 2021. "Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights". *arXiv: Learning*.

[28] Morris, B., 2018. Mastering the learning rate to speed up deep learning. `https://brandonmorris.dev/2018/06/24/mastering-the-learning-rate/`. Accessed: 2022-06-30.

[29] Smith, L. N., 2018. "A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay". *ArXiv,* **abs/1803.09820**.

[30] Gessert, N., Nielsen, M., Shaikh, M., Werner, R., and Schlaefer, A., 2020. "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data". *MethodsX,* **7**.

[31] Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. V., 2019. "Autoaugment: Learning augmentation strategies from data". *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123.

[32] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V., 2020. "Randaugment: Practical automated data augmentation with a reduced search space". *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3008–3017.

[33] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B., 2020. "Augmix: A simple data processing method to improve robustness and uncertainty". *ArXiv,* **abs/1912.02781**.

[34] Ha, Q., Liu, B., and Liu, F., 2020. "Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge". *ArXiv,* **abs/2010.05351**.

[35] Ali, D. A.-R., Li, J., and O'Shea, S. J., 2020. "Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images". *PLoS ONE,* **15**.

[36] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E., 2020. "A simple framework for contrastive learning of visual representations". *ArXiv,* **abs/2002.05709**.

[37] Devries, T., and Taylor, G. W., 2017. "Improved regularization of convolutional neural networks with cutout". *ArXiv,* **abs/1708.04552**.

[38] Archive, I., 2019. Evaluation score. `https://challenge.isic-archive.com/landing/2019/`. Accessed: 2022-06-30.

[39] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., 2019. "Pytorch: An imperative style, high-performance deep learning library". In NeurIPS.

[40] Wightman, R., 2019. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`.

[41] Falcon, W., 2022. Pytorch lightning. `https://www.pytorchlightning.ai/`. Accessed: 2022-07-30.

[42] Biewald, L., 2022. Weights biases. `https://wandb.ai/`. Accessed: 2022-07-30.

[43] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A., 2017. "Inception-v4, inception-resnet and the impact of residual connections on learning". In AAAI.

[44] Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K., 2017. "Aggregated residual transformations for deep neural networks". *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995.

[45] Bao, H., Dong, L., and Wei, F., 2022. "Beit: Bert pre-training of image transformers". *ArXiv,* **abs/2106.08254**.

[46] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., 2021. "Swin transformer: Hierarchical vision transformer using shifted windows". *2021 IEEE/CVF International Conference on Computer Vi-*

sion (ICCV), pp. 9992–10002.

[47] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B., 2022. "Swin transformer v2: Scaling up capacity and resolution". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11999–12009.

[48] Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. K., 2019. "Billion-scale semi-supervised learning for image classification". *ArXiv,* **abs/1905.00546**.

[49] Loshchilov, I., and Hutter, F., 2017. "Sgdr: Stochastic gradient descent with warm restarts". *arXiv: Learning*.

[50] Smith, L. N., 2017. "Cyclical learning rates for training neural networks". *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472.

[51] Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., and Ridella, S., 2012. "The 'k' in k-fold cross validation". In ESANN.

[52] Kasmi, R., and Mokrani, K., 2016. "Classification of malignant melanoma and benign skin lesions: implementation of automatic abcd rule". *IET Image Process.,* **10**, pp. 448–455.

[53] Yuan, L., Hou, Q., Jiang, Z., Feng, J., and Yan, S., 2022. "Volo: Vision outlooker for visual recognition". *IEEE transactions on pattern analysis and machine intelligence,* **PP**.

[54] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and J'egou, H., 2021. "Going deeper with image transformers". *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 32–42.

[55] Yang, J., Sun, X., Liang, J., and Rosin, P. L., 2018. "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria". *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1258–1266.