Application of Deep Learning general-purpose neural architectures based on Vision Transformers for ISIC melanoma classification

Author
David Dueñas Gaviria

Universitat Politécnica de Catalunya (UPC) · BarcelonaTech
Universitat de Barcelona (UB)
Universidat Rovira i Virgili (URV)

Supervisor and Co-supervisor
Dr. Petia Ivanova Radeva · Universitat de Barcelona
Dr. Md Mostafa Kamal Sarker · University of Oxford

A thesis submitted for the degree of

Master's degree in Artificial Intelligence

Barcelona, October 18, 2022

This thesis is submitted to the Facultat D'informàtica de Barcelona, Universitat Politècnica de Catalunya in fulfilment of the requirements for the master's degree in artificial intelligence.

October 18, 2022

Copyright © 2022
David Dueñas Gaviria

Acknowledgements

I would like to thank my family and friends for supporting me throughout my academic journey. To my professors, impeccable in their lessons and their disposition for teaching and reciprocal learning. To my supervisor Dr. Petia Radeva, and co-supervisor Dr. Md Mostafa Kamal who with their instructions and constant support have nurtured my professional development and encouraged me to go further, thank you.

"If I have seen further than others, it is by standing upon the shoulders of giants."

Isaac Newton

Abstract

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/).

Contents

List of Figures

Chapter 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 WHO (2017), Forsea (2020). According to a study by Arnold et al. (2022) 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 WHO, 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 Leonardi et al. (2018), and yet, within a melanoma diagnosis, the 5-year survival rate exceeds 90% ACS (2022); 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 performance and assessing it particularly on the melanoma.

On the whole, it is well established that morbidity and death from melanoma can be dramatically decreased by early diagnosis, and the stage of the malignant lesion—intrinsically related with its depth and time of growth—determines the likelihood that the patient will die from the condition Tejera-Vaquerizo et al. (2012). As a result, early detection is critical, and current methods include patient education, periodic self-examination of the skin by patients, and full-body skin exams by medical experts.

As medical professionals' and patients' needs for technology have increased, so have the demands for automated skin cancer diagnosis Chang et al. (2013); 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 Belilovsky et al. (2019). 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) Russakovsky et al. (2015). However, there are a number of other DL techniques that have been gaining in popularity in recent years. Particularly Vision Transformers (ViTs) Dosovitskiy et al. (2021), 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 Vaswani et al. (2017). Although considerable research has been done on the use of ViTs for medical image classification, see Chen et al. (2021), 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.

Based on the ISIC 2019 Challenge for the classification of skin lesions and recent publications, this study assesses the state of the art in the categorization of dermoscopic skin disease images. Since medical image datasets frequently exhibit a class imbalance, a number of techniques for deep NNs training on unbalanced datasets have been examined. As deep NNs require a large amount of training data, other publically accessible dermoscopic and clinical imaging datasets of skin lesions have been considered for supplementing the ISIC 2019 training data.

Along with that, 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 ensemble 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.

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 thesis 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 a body of literature that was consulted and served as the foundation for the proposed solution. The third chapter shows the deep learning fundamentals elements to came up with the solution, while the forth chapter goes over our model description and implementation processes training the various ViTs and CNNs models used and generating predictions. The fifth chapter discusses everything about the data, including who hosts it, what it consists of, and its relations and properties. The sixth chapter displays and summarizes all of the results acquired along with their analysis, and the seventh chapter encompasses 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.

Code is publicly available under MIT License at:
`https://github.com/blobquiet/SIIM-ISIC-Melanoma-Classification`

Chapter 2

Literature Review

2.1 The problem of automatic analysis of dermoscopic images

2.1.1 Dermoscopy Images

Dermoscopy is the go-to approach employed in skin lesion datasets; it generates high-resolution images, allowing for many versions of the same lesion, and eliminates surface reflection of skin, improving visibility of deeper layers of skin, making it a standard technique for skin lesion imaging Sun et al. (2016). Prior research has shown that when used by dermatologists, dermoscopy provides improved diagnostic accuracy, in comparison to standard photography Celebi (2021). As such, it is the imaging technique of choice for the ISIC skin image challenges. Figure 2.1 depicts a cancerous melanoma imaged with dermoscopy.

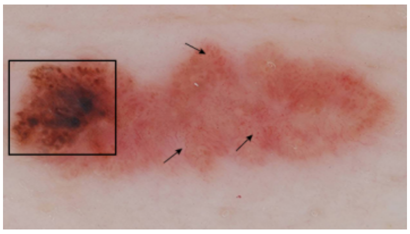


Figure 2.1: Dermoscopy melanoma image. Commonly a non-symmetric halves lesion, with border irregularity. it has uneven pigmented colours, which tent to grow in size and diameter (American Cancer Society)

The use of dermoscopy images has demonstrated great potential in the classification of skin diseases Chang et al. (2013), Sun et al. (2016). However, dermoscopy has limitations in that it requires specialized medical equipment and specialist skills Kasmi and Mokrani (2016), which is why datasets relating to skin images are still being gathered with clinical images ISIC-Colaboration. (2020).

2.1.2 Clinical Skin Image Recognition

Clinical skin images are created under various lighting conditions, and lesions have non-uniform focal lengths because they are taken from easily accessible devices, such as smartphones, Yang et al. (2018). This enables them to have far more disease categories in clinical images than dermoscopic images, as well as a much larger dataset size and accessibility in comparison to dermoscopic images. Figure 2.2 illustrates a clinical skin image of a malignant melanoma obtained from The University of Iowa's Department of Dermatology.

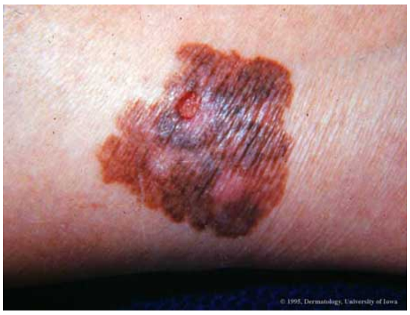


Figure 2.2: Clinical skin image of a malignant melanoma from the University of Iowa (of Iowa Health Care (2020))

The classification of clinical skin disease images has been a continuing activity over the past decade Glaister (2013), Alquran et al. (2017), Narayanan (2020), Razeghi et al.. The majority of them concentrated on Machine Learning (ML) methods like K-nearest-neighbor (KNN), Support Vector Machine (SVM), or Random Forest (RF), which largely relied on a human-in-the-loop and expert knowledge and were far from automated. However, some studies, such as Esteva et al. (2017), yu Zhu et al. (2021), Wu et al. (2020), have been motivated to use clinical images from either public or private datasets with the goal of a data-driven approach that employs DL techniques to address the skin lesion recognition task and automate the process using a large body of images, which are more accessible than dermoscopy.

2.2 Related Work

2.2.1 Skin Lesion Computer-Aided Diagnosis

2.2.1.1 Computer-Aided Diagnosis of Skin Lesions using Conventional Digital Photography: a Reliability and Feasibility Study

To categorize skin lesions with melanocytic and non-melanocytic conditions using conventional digital macrographs from an electron microscope, researchers from Kaohsiung Medical University have created a Computer-Aided Diagnosis Software (CADx) Chang et al. (2013). Feature extraction and SVM analysis were carried out using ML approaches, and this yielded useful information, particularly through the color correlation and its Principal Component Analysis (PCA). The system performance metrics were as follow: ROC AUC of 0.949, sensitivity and specificity of 85.63% and 87.65% respectively. In contrast, those of the dermatologist: sensitivity, specificity, and accuracy were 83.33%, 85.88%, and 85.31%, respectively. Overall, with comparable results to that of the clinicians at their institute it was noticed that since 2013 the growing interest and the limitations in accessibility for non-melanocytic data were being recognized.

2.2.1.2 Dermatologist–level Classification of Skin Cancer with Deep Neural Networks

In this study, which was published as Esteva et al. (2017), the authors use a dataset of 129,450 clinical photos—including 3,374 images of dermoscopy—labeled by dermatologists, representing one of the largest datasets used to date (2017) Masood and Al-Jumaily (2013), to demonstrate the ability of CNNs for generalized classification. They outline the difficulties that arise from the use of non-standardized equipment and its variability in factors like zoom, angle, and lighting, which significantly affect classification in smartphone-like real-world photographs in contrast to the more standard instruments used in dermoscopy and histological images obtained through invasive biopsy and microscopy. With their Computer-Aided system, they concentrated on avoiding the extraction of visual characteristics particular to a certain domain, as well as time-consuming preprocessing, lesion segmentation, and other labor-intensive classification requirements. Using a single, previously trained CNNs for both photographic and dermoscopic purposes, the system proposed contrasts this by demonstrating the viability of a data-driven method. Its performance was then compared to the ground truth of 21 certified dermatologists on biologically verified clinical images with two cases: the identification of the two most common cancers, keratinocyte carcinomas and benign seborrheic keratoses, and for the second malignant melanomas againts benign nevi which represents the identification of the deadliest skin cancer, once again with comparable results.

2.2.2 Data Augmentation for Skin Lesion Analysis

The influence of thirteen data augmentation scenarios was examined in the study by Perez et al. (2018) for the classification of melanoma trained on three robust CNNs (Inception-v4, ResNet, and DenseNet). Traditional color and geometric transformations are included in the scenarios, along with more unique augmentations including elastic transforms, random erasing, and a new augmentation that mix the skin lesions. They also examine the application of data augmentation during testing and its effects on different dataset sizes. The findings of this study demonstrate the benefits of augmentation in training data as well as the influence of augmentation in test data on the detection of melanoma. Nonetheless, the possibility of fine-tuning hyperparameters and exploring adjustments to geometry and color still remain unexplored. Finally, model ensembling was determined to yield significant improvements. Mix augmentation has been shown to provide less favorable results in this task, and manual augmentation processing techniques were concluded to be useful but should be used with caution because they could produce unreliable images.

2.2.3 The ISIC challenge

2.2.3.1 Skin Lesion Classification Using Ensembles of Multi-Resolution EfficientNets with Meta Data

The work in Gessert et al. (2020) outlines the approach that won the ISIC 2019 Skin Lesion Classification Challenge's first ranking for both assignments. They use a loss balancing strategy to solve the common issue of class imbalance in the classification of these skin lesions. They focused on a variety of EfficientNets with different input cropping techniques and input resolutions to deal with the image resolutions. Likewise, for the test set unknown class they used a data-driven approach of images of healthy skin along with a two-path design that merges dermoscopic and meta data into the model. They conclude that predicting an unidentified class and making the best use of meta data remain difficult issues. Their best ensemble achieved a sensitivity of 74.2% using five-fold crossvalidation. Overall, on the official test set the method is ranked first for both tasks with a balanced accuracy of 0.636 for task 1 and 0.634 for task 2, with an ensembling strategy where they searched for the optimal subset of models.

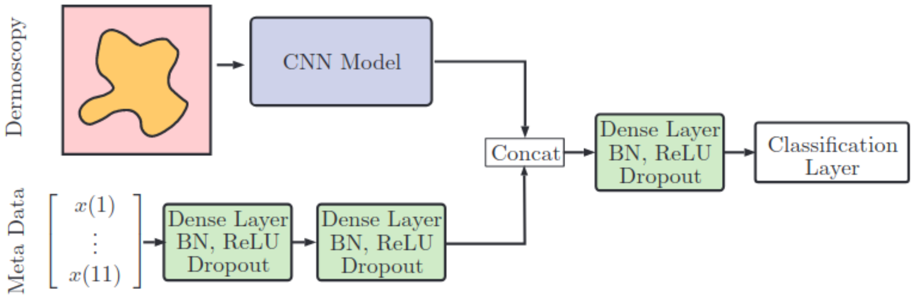

Dermoscopy
CNN Model
Concat
Dense Layer BN, ReLU Dropout
Classification Layer
Meta Data
$x(1)$
⋮
$x(11)$
Dense Layer BN, ReLU Dropout
Dense Layer BN, ReLU Dropout

Figure 2.3: Pipeline for Gessert et al. (2020), combining dermoscopic image processing and meta data processing

The pipeline of this work can be seen in Figure 2.3. This work is relevant, because it uses an EfficientNet with a preprocessing technique consisting of a cropping strategy, which was validated in this study. The EfficientNets have been recurrent since 2019, when they first emerged in the research introduced by Tan and Le (2019), and they are typically the most commonly utilized CNNs for such challenges since their appearance. Finally, one particular method used here was the augmentation technique, which employed a unconventional augmentation with a CutOut approach for regularization Devries and Taylor (2017) and a color augmentation called 'Shades of Gray and Color Constancy' Finlayson and Trezzi (2004), which they think to be significant to their outcome.

2.2.3.2 Identifying Melanoma Images using EfficientNet Ensemble Winning Solution to the SIIM-ISIC Melanoma Classification Challenge

At the ISIC 2020 Skin Lesion Classification Challenge, The solution Ha et al. (2020), came in first place for melanoma classification. It was composed of a large number of CNNs, including EfficientNet B3-B7, SE-Resnext101, and ResNeSt101. Input sizes range from 384 to 896, with the majority being images and some containing patient and image-level metadata. They have consolidated on a validation strategy based on 5-fold cross validation (CV) on the combined data from 2018, 2019 and 2020 challenge, a target softmax probability for melanoma prediction and trusting their AUC metric instead of the leaderboard public score. Finally, their best ensemble of the winning contribution had an AUC of 0.96, while the private leaderboard had an AUC of 0.9490 reaching first rank. The pipeline can be found in Figure 2.4.

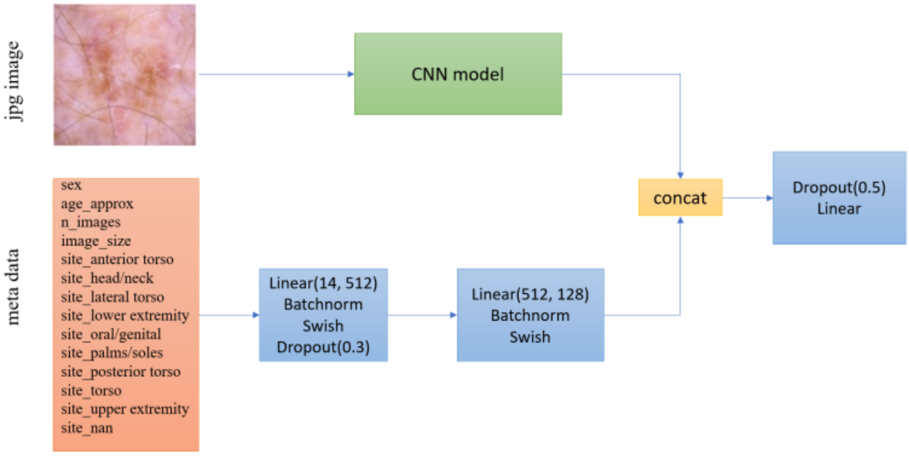


Figure 2.4: Pipeline for Gessert et al. (2020), combining dermoscopic image processing and meta data processing

The importance of this work consists of using the nine class diagnosis from previous years, and after the training process focusing solely on the prediction of the melanoma diagnosis, rather than the binary benign or malignant task. They demonstrated that a diagnosis of skin lesions provides more details than a binary target for malignant melanoma. Nonetheless, they describe a somewhat brute force method that involved training a range of 18 models under various settings, including models with and without metadata and at various resolutions. Finally, it is important to keep in mind that the augmentation technique employed is based on computational validation, and certain images have been transformed at the point of deformity, which may be an area that needs improvement in order to have more reliable predictions.

2.2.3.3 Analysis of Skin Lesion Images with Deep Learning

The work made by Steppan and Hanke (2021), evaluated the state of the art CNNs in the classification of ISIC 2019 Challenge. A combination of EfficientNet architectures, along with ResNext and Inception architectures pre-trained on the ImageNet was trained on a crafted dataset comprised of dermoscopic and clinical images of skin lesions, which demonstrated improvements in outlier detection. Nonetheless, the training method was proven computationally expensive, using both transfer learning in several phases, with freezed layers and model finetuning on a low learning rate (LR) for several epochs. Overall balanced accuracy was further improved by using an ensemble of only four independently trained models.

This method's relevance can be attributed to the data-driven approach it adopts to handle outliers and the variety of publicly accessible datasets it uses. Even though this method did not win the 2019 ISIC challenge and did not use metadata, the utility of threshold shifting for handling the imbalanced problem was examined by reviewing undersampling/oversampling, balanced cross entropy, and finally thresholding, which focuses on multiplying the model predictions by inverse class frequencies to approximate actual probability distributions.

Chapter 3

Deep Learning Fundamentals

DL refers to a subset of ML which, subsequently refers to a subset of Artificial Intelligence (AI) Schmidhuber (2015). DL has gained popularity in the computer vision field in the recent decade since the appearance of a multitude of factors: namely powerful Graphic Processor Units (GPUs) capable of parallel processing of huge datasets of images Steinkrau et al. (2005). Before DL was a popular topic, Artificial Neural Netowrks (ANNs) were amply discussed in the 1960's, with the Shallow NNs models Schmidhuber (2015) which later gave birth to DL Hinton and Salakhutdinov (2006), and important concepts like gradient descent and back propagation developed from the 60's and 70's to the 80's onwards Belilovsky et al. (2019). Figure 3.1 shows the AI taxonomy described above.

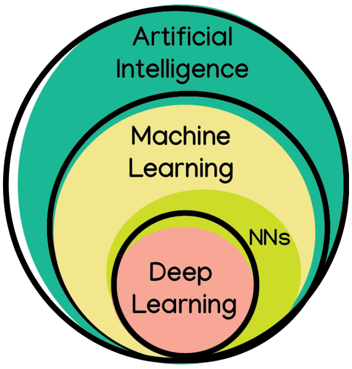


Figure 3.1: AI taxonomy

NNs were in essence developed from the domains of psychology and neurophysiology rather than computer science Pircher et al. (2021), and they are explicitly modeled by a biological architecture. Despite being a mathematical abstraction of biological neurons, behavior that resembles them emerges when exposed to enough data and large layers of units Rosenblatt (1958).

Units in NNs have an activation that represents a linear combination of the neuron's input and its parameters. The total input x_j from the unit j is a linear function from the output y_j Litjens et al.; w_{ij} and b_j are the sets of weights and biases that comprise the equation 3.1:

$$x_j = b_j + \sum (y_w w_{ij}) \tag{3.1}$$

The Multilayer Perceptron (MLP) is the most conventional and extensively used type of NNs. The several hidden layers of an MLP make it possible to construct deep NNs, and the non-linear activation functions are precisely where the MLP value resides. Typically, it uses the logistic or sigmoid function Rumelhart et al., shown in 3.2:

$$logistic(x_j) = \frac{1}{1 + e^{-x_j}} \tag{3.2}$$

Likewise, through the use of a softmax function, the final layer activations of the network are converted to a probability distribution P_j over all classes k:

$$p_j = \frac{e^{x_j}}{\sum_k e^{x_k}} \tag{3.3}$$

Overall, with the appearance of deep learning architectures, some of the most popular models integrating low level characteristics such as edges and more complex forms with semantic meaning have thrived. CNNs are now the most extensively employed in (medical) image processing, however another paradigm called Vision Transformers, emphasizing on self-attention is gaining popularity and beating a broad range of visual benchmarks Bai et al. (2021).

3.1 Transfer Learning

In some scenarios, it can be difficult and computationally costly to obtain a considerably large amount of training data that fits the feature space and allows to build a DL learner from scratch that could be able to generalize on the test data Weiss et al. (2016). Therefore, in order to create a high-performance learner for the target domain, it is necessary to use a large amount of information from pre-trained models in the ImageNet, which corresponds to a dataset of 1000 classes and over one million images, to transfer the knowledge gathered in the form of low level visual features (as blobs, edges, shapes, textures, and colors) and further train on different images from a new dataset. In the context of ML, transfer learning is a strategy that focuses on the retention of knowledge acquired from one task and applies it to perform in a similar, but different task Pan and Yang (2010). Therefore, when applying transfer learning, a model that has previously been trained on a large dataset is adjusted to specific data, which is typically of a smaller population that would be unsuitable to build a deep learning model from scratch Krizhevsky et al. (2012).

Fine-tuning, on the other hand, corresponds to a variant of transfer learning in which not only fully-connected layers, but also a greater number of layers become trainable, and these are often adjusted with a lower learning rate to gradually improve network performance on the new dataset Vrbancic and Podgorelec (2020). Architecture upgrades could include increasing the number of additional trainable parameters, as well as freezing and unfreezing layers for a short number of epochs. In this manner, only specific knowledge mined from the previous task is maintained, while trainable parameters from the network's final layers—which generally carry semantic information—are renewed.

3.2 Convolutional Neural Networks

CNNs are a subset of NNs that have excelled at image recognition and classification Anderson et al. (2018), among other tasks. CNNs drew some indirect inspiration from nature as well. They are based on the neocognitron Fukushima (2004), which was created with the goal of simulating the behaviors of cells in the visual cortex of cats and monkeys. They are composed of many layers, each of which is able to learn to recognize patterns that likewise recognize features in images, which is why they are often used for tasks such as object detection and facial recognition Anderson et al. (2018). The primary application of CNNs is in pattern recognition, with images as input and the layer parameters revolve around the use of learnable kernels preceded by subsampling or pooling layers O'Shea and Nash (2015), with the goal of reducing dimensionality and, as a result, the computational complexity of the model. The common architecture of CNNs can be seen in Figure 3.2.

Overall, convolutional layers which are stacked on top of one another, are the foundation of a CNN's architecture. These layers are followed by sub-sampling (pooling) layers, which are repeated several times until they reach the fully connected layer, which is fed forward before the final layer that predicts the output.

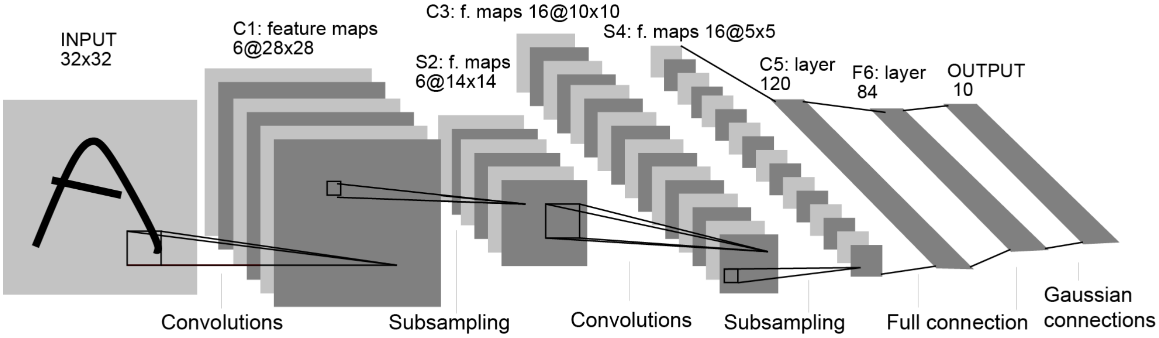


Figure 3.2: Typical architecture of a CNNs (LeNet-5 LeCun et al. (1998))

The input of each layer is then organized into three dimensions: spatial dimension (height and width) and channel count (depth).

3.2.1 EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks Student

The compound scaling approach, which is founded on the assumption that multiple scaling dimensions are not independent, lies at the heart of EfficientNets, as depicted by Tan and Le (2019).

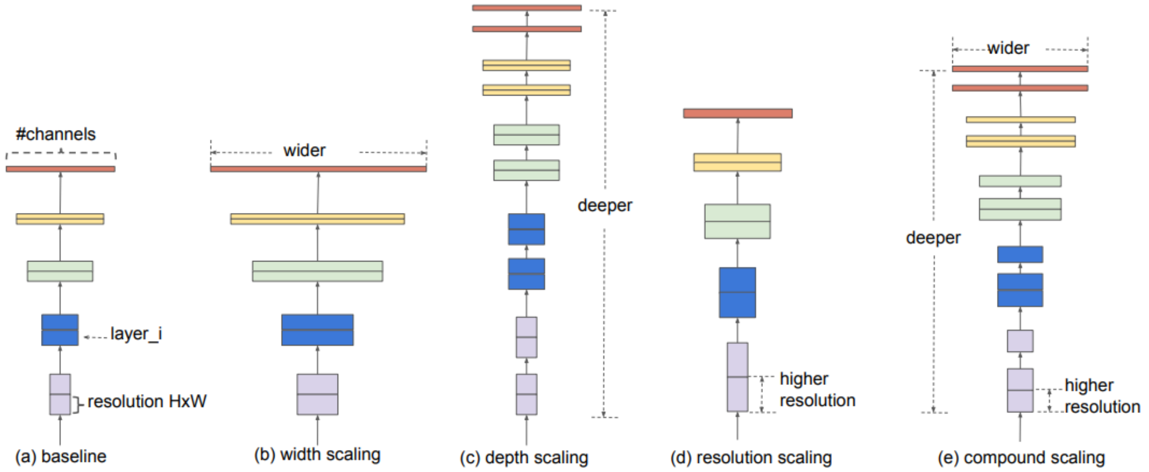


Figure 3.3: Re-scaling process for EfficientNets. (a) depicts a baseline network. (b) A feature map scaled by width. (c) Depth, which adds more layers. (d) Using a greater resolution, and (e) A mix of the previously described. (Tan and Le (2019))

There are plenty arbitrary methods for scaling up CNNs. Figure 3.3 depicts the model's various scaling options for CNNs. One method is to add additional feature maps to each layer, resulting in more layers and greater resolution; however, by scaling them with a constant ratio, B0 to B7 EfficientNets emerge. The EfficientNets main idea is that the better the image resolution from the input, the more layers the network requires to enhance the receptive field and more channels to catch more patterns in the image. Finally, the work given in Xie et al. (2020) increases EfficientNets generalization performance by combining a teacher-student strategy with a semi-supervised technique that takes advantage of massive amounts of unlabeled data. With it, an EfficientNet model is trained in labeled data to produce pseudolabels, and then a bigger EfficientNet is trained as the student on these new datasets; this process is repeated by including noise in the student learning process with both networks switching roles interatively.

3.2.2 ConvNeXt: A ConvNet for the 2020s

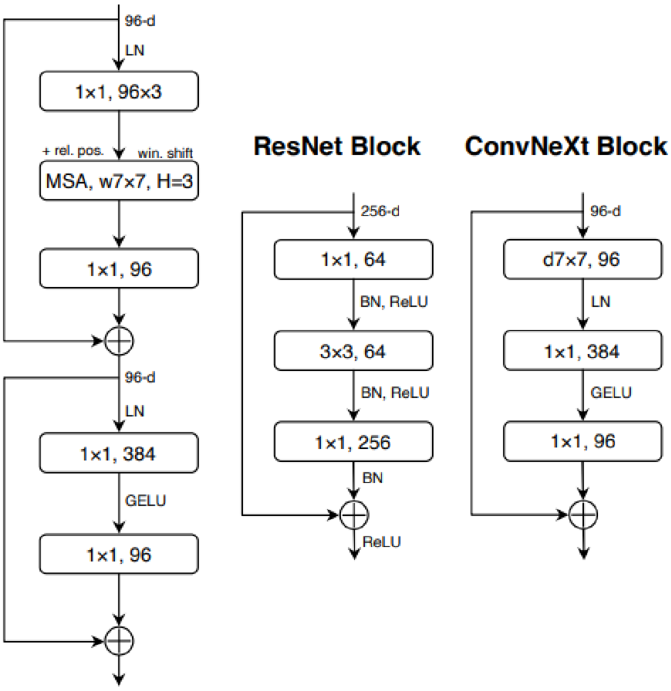


Figure 3.4: (Liu et al. (2022b))

Motivated by the architectural design of Swin Transformers, the work in Liu et al. (2022b) has designed a road-map in order to combine the global semantics provided by ViTs and the locality advantages of ConvNets. Moreover, the improvements which began with taking a ResNet-50 He et al. (2016), aims to combine the global semantics provided by ViTs and the locality advantages of ConvNets. Improving on with similar training techniques as ViTs; meaning using AdamW optimizer, more epochs, heavy data augmentation and regularization. Then, increasing the kernel and addding fewer activation functions and normalization layers to end up with a block design that had demonstrated to benefit the CovNeXt outperforming the Swin Transformer. Figure 3.4 illustrates a Swin Transformer block along with the ResNet and ConvNeXt block. It is worth noting that Swin Transformer block is the most sophisticated due to the presence of multiple specialized modules and two residual connections.

Finally, the depthwise convolutions from the ResNeXt block, which are analogous to the sum-operation in self-attention, plus the network widening result in an improvement that may be further stretched by adding a series of ViTs adjustments. Specifically, increasing the kernel size allows the CovNeXt to outperform the Swin Transformer by changing the activation function and replacing Batch Normalization with Layer normalization.

3.3 Vision Transformers

The introduction of transformers by Vaswani et al. (2017) has significantly enhanced the discipline of Natural Language Processing (NLP). With the addition of the self-attention module, transformer can perform at the cutting edge on a variety of NLP tasks by efficiently capturing the non-local interactions between all input sequences. ViTs appeared as a proof that reliance on CNNs is not always necessary and a pure transformer applied directly to sequences of image patches can be a promising visual recognition method to consider Dosovitskiy et al. (2021). Recent research also claims that ViTs are substantially more resilient than CNNs in addition to displaying competitive performance on a variety of visual benchmarks.

A typical ViT architecture can be found in Figure 3.5 and work by dividing the image into fixed-sized patches that are linearly projected and given positional embeddings, ViTs hence focuses on global attention over pixels. The output of the standard transformer encoder, which receives these learnable embedding as input, is added to the sequence through the classification head to provide predictions. Furthermore, the most important component of ViTs is its Multi-Head attention module, originally from Vaswani et al. (2017), which runs numerous separate attention outputs in parallel until they are concatenated and linearly translated into the desired dimension, assisting the model in concentrating on the important information inside each patch embedding.

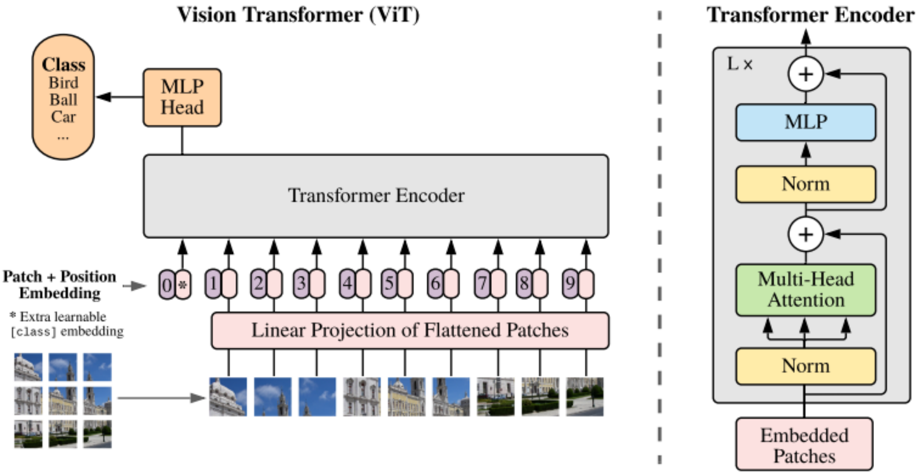


Figure 3.5: Typical ViTs architecture (ViT), Dosovitskiy et al. (2021). ViTs focuses on global attention over pixels by splitting the image into fixed-sized patches that are linearly projected and assigned a positional embedding. These learnable embedding are then fed into a standard Transformer Encoder, whose output is added to the sequence via the classification head to provide predictions .

3.3.1 Data-efficient Image Transformer (DeiT)

DeiT has been recognized as one of the robust ViTs, as authors Touvron et al. (2021a) have shown it clearly outperforms ViT and EfficientNets by a margin. However, not only do the authors propose distillation to train transformers, but they also provide insights on how to make the training more efficient with far less data and far less compute power for a high performance classification model. Figure 3.6 depicts DeiT's main proposition. The distilation token works in the same way as a class token and new distillation embedding allows the model to learn from the teacher's output, remaining complementary to the class embedding while distilling. The teacher is supposed to be a powerful image classifier, and learning happens as a result of the trade-off between accuracy and image throughput.

3.4 Ensemble Learning

Deep ensemble can soon be understood with a probability principle known as Condorcet's Jury Theorem Boland (1989), which states that if multiple jurors have a probability greater than 50% of arriving at a true verdict based on independent decisions, then the probability of a true verdict on a group of jurors can be guaranteed to be close to 100 percent as the number of jurors in the group increases. DL approaches frequently have constraints, and a single DL model will not always

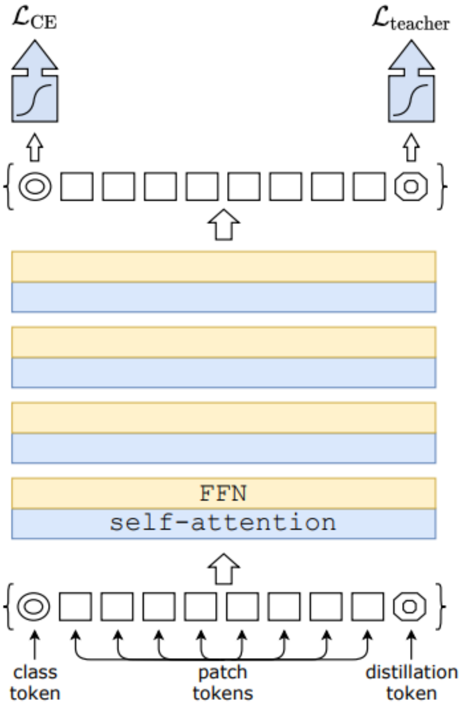


Figure 3.6: DeiT distillation procedure: by means of the self-attention layers, a new distillation token interacts with the class and patch tokens. This distillation token is used to replicate the (hard) label predicted by the instructor rather than the genuine label on network output. (Touvron et al. (2021a))

achieve top performance in the real-world Rokach (2009). Moreover, research from Breiman (2004) and Freund and Schapire (1999), has shown that an ensemble of many models is frequently more accurate than any single classifier, and assembling numerous algorithms may be accomplished by aggregating predictions either by a voting scheme or more commonly via averaging, both of which are addressed in this study. Overall, ensembles are expected to increase performance by reducing a deep learning model's large variance via training many models on diverse initializations. The majority voting scheme works by the Equation 3.4, where the class label $\hat{y}$ is determined by the majority of the votes of each classifier C_m.

$$\hat{y} = argmax\,\{C1(x), C2(x), .., C(x)_m\} \tag{3.4}$$

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{3.5}$$

Averaging, on the other hand, goes by the equation 3.5. Where $\bar{p}$ represents the mean value of an ensemble prediction, N the total number of models and p_i the variable predictions of the models.

3.5 Out of Distribution

Modern NNs are known to generalize well when training and testing data are sampled from the same distribution Liang et al. (2018). However, this is not always the case in practice because control over the testing data distribution is not guaranteed Sastry and Oore (2019). In such scenarios, a DL classifier based on NN is unlikely to be a credible source since it would give an overconfident and incorrect prediction on such data. This data is referred to as Out-of-Distribution (OOD), and detecting it remains a barrier in the medical field Berger et al. (2021). In the work presented by Pacheco et al. (2020), a study was carried out for deep neural classifiers on the 2019 ISIC dataset. In particular, the Gram-OOD matrix was examined for OOD detection, and they present an extended Gram-OOD* that performs similar to the data-driven approaches from the submissions made for the 2019 challenge:

$$G_l^p = F_l G_l^{p\top}, \tag{3.6}$$

$$\Delta(\breve{x}) = \sum \frac{\delta_l(\breve{x}_c)}{E_{Va}[\delta_l]}.$$

To highlight prominent features, Equation (3.6) computes high order Gram-Matrices of order p with Fl corresponding to activations at layer l. The first step is to compute pairwise correlation between the obtained feature maps, both in convolutional layers and activation layers; next, the layerwise deviations from the gram matrices are computed so that it is possible to know how much a sample deviates from the max/min values over the training data. Finally, the original method computes the total deviation by summing its layerwise deviation across all layers. In Equation (3.7), the expected deviation from the validation data $E_{Va}[\delta_l]$, is computed using the validation set, which is chosen as a subset of the training data, avoiding the need for OOD datasets, in contrast to techniques such as Liang et al. (2018), which need both In-distribution and Out-of-distribution datasets.

$$\tilde{G_b^p} = \frac{\hat{G_b^p} - min(\hat{G_b^p})}{max(\hat{G_b^p}) - min(\hat{G_b^p})} \tag{3.7}$$

The enhanced Gram-OOD*, shown in Figure 3.7, adds an extra normalizing layer between the pairwise correlations and the layerwise variances. The normalization procedure is depicted in equation (3.7). Finally, in our work both a

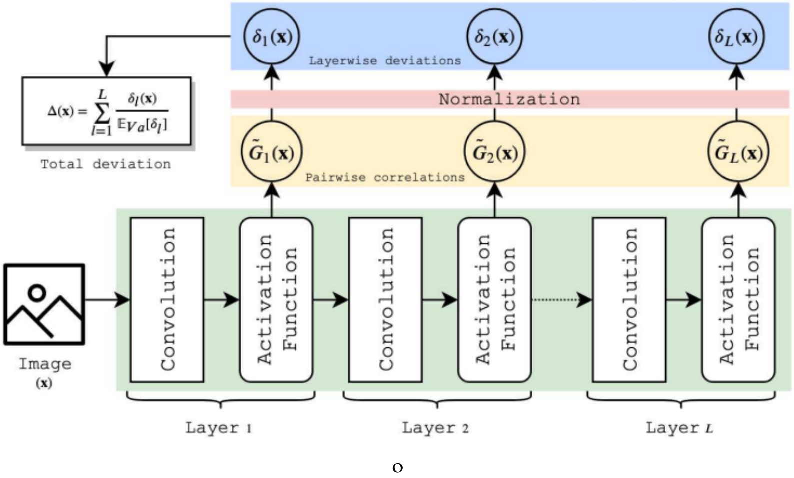


Figure 3.7: Gram-OOD* overview Pacheco et al. (2020). It depicts the pipeline of the original Gram-OOD with the addition of the normalization layer in between the pairwise correlation and the layerwise deviations computation.

data-driven strategy based on external data for the unknown class in the ISIC 2019 Challenge and an adaptation from the Gram-OOD* were applied.

3.6 Super-Convergence

In Smith and Topin (2019) research, the authors study a phenomenon known as ”super-convergence”, where it is demonstrated that NNs can be trained an order of magnitude quicker than with regular training techniques. In super-convergence, networks are trained with high LRs in an order of magnitude fewer iterations and with better final test accuracy than when a constant training regime is used. Training with a single LR cycle and a high maximum LR is a critical component of super-convergence. Figure 3.8 depicts the test accuracy comparison on the left, of the ResNet-56 using the Cifar10 dataset, with the blue line representing the standard training regime and the orange line representing the LR cycle. The right depicts the key elements of super-convergence: a minimum LR and a peak with a maximum LR.

A key finding that enables super-convergence training is that high LRs regularize the training, requiring a decrease in all other kinds of regularization in order to maintain an appropriate regularization balance. Experiments also show that super-convergence is conceivable with a wide range of datasets and architectures, as long as the regularization impacts of high LRs are balanced by minimizing other

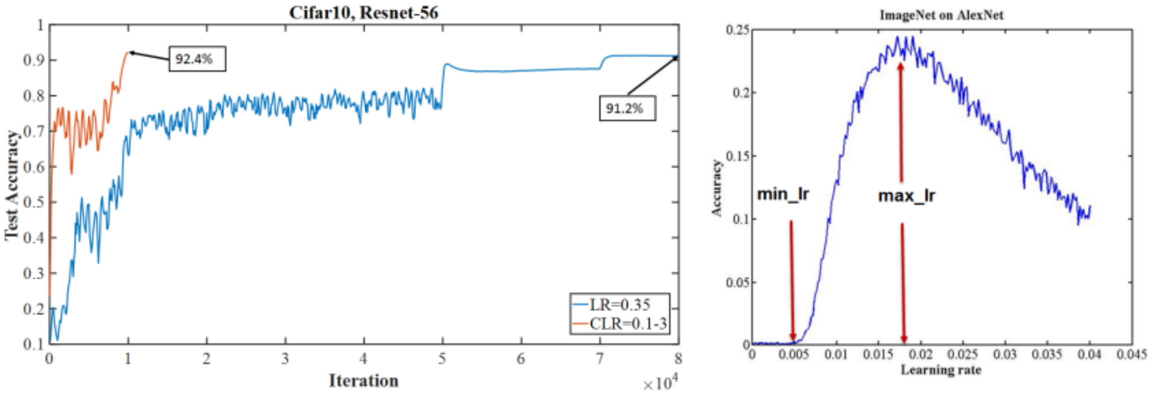


Figure 3.8: Example of a super-convergence image and a single LR cycle; the left represents test accuracies when compared to a standard training regimen, and the right displays the two basic components of super-convergence, with a peak on the max lr and a predefined min lr.

types of regularization. Furthermore, in Smith (2018) they showed that when the amount of labeled training data is restricted, super-convergence delivers a higher performance gain than normal training. Overall, the extension the authors proposed was to use a cycle that is always less than the sum of the iterations/epochs, and then, during the subsequent iterations, the learning rate might drop by a factor of many orders of magnitude less than it did at first.

3.7 Data augmentation

Data augmentation approaches are frequently used to train advanced applications with a limited amount of images. Data augmentation in the DL field encompasses a wide range of techniques used to generate training samples taken from the original set by applying various random perturbations without changing the class labels. Image augmentation techniques refer to transforming input images into fresh yet realistic samples of new images that will help the network generalization. Although current generative models demonstrate that realistic medical image generation using more complex techniques is conceivable Galdran et al. (2017) Costa et al. (2017), semantic-preserving transformation, such as brightness and contrast, shift and scale, flip and rotation are often performed with promising results. See Figure 3.9 for semantic-preserving reference Shorten and Khoshgoftaar (2019). Nonetheless, incorporating domain-specific knowledge about the expected images remains open, which can help create a richer set of new images while preserving consistency.

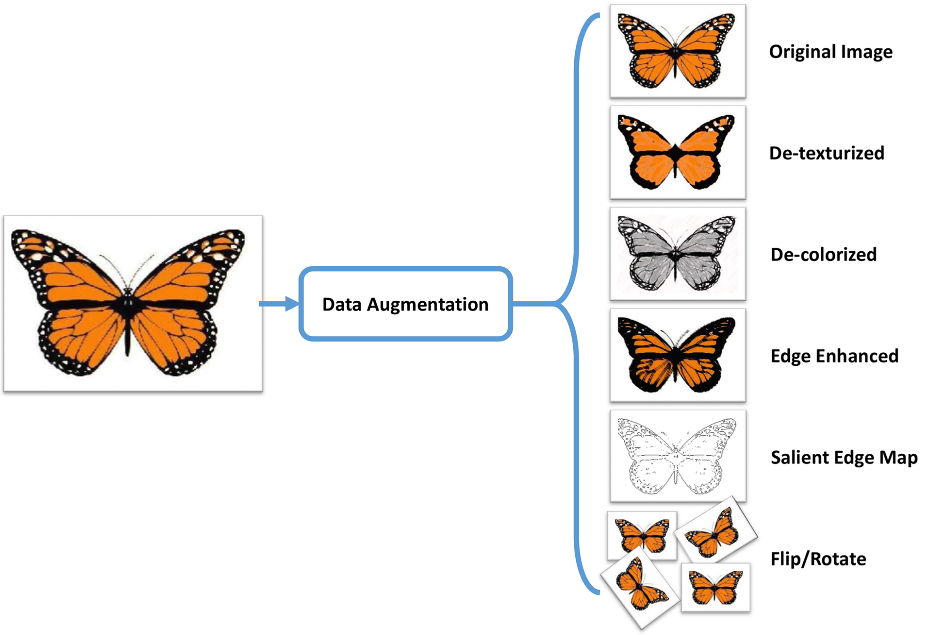


Figure 3.9: Example of data augmentation preserving semantic transformation (Ahmad et al. (2017)). Data augmentation generally entails crops, rotations, translations, scaling, mirroring, etc.

3.8 Test Time Augmentation

The use of test-time data augmentation for the assessment of deep NNs was first shown in the paper cited by Ayhan and Berens (2018). Since then, Test Time Augmentation (TTA) has gained popularity as a computer vision technique that, regardless of the underlying model, uses data augmentation on the inference stage to increase model accuracy and decrease generalization errors Shanmugam et al. (2020). Figure 3.10 illustrates the concept of TTA; when using TTA, inference is carried out on various augmented variations on each one of the test images (such as random crops, flips, color-contrast), and the predictions are then averaged to improve the performance of the models.

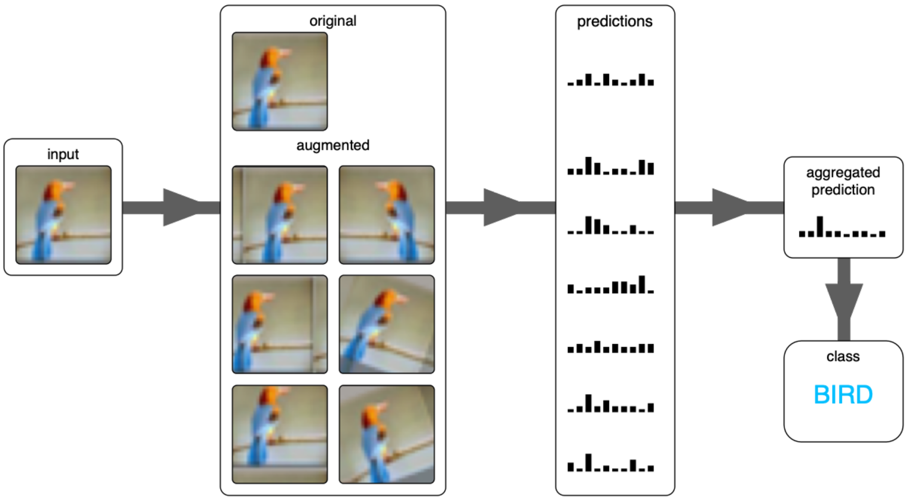


Figure 3.10: Example of TTA for boosting model performance at inference Dufour (2020). The aggregation of predictions occurrs after multiple versions of the same test image are generated.

Chapter 4

Our Proposal: A New Method for Skin Lesion Classification

In this chapter, we will 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:

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

4.1 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) Touvron et al. (2021a), 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 Tan and Le (2019), trained on Noisy-Student weights Xie et al. (2020) and using a scaling technique to equally scale the network's width, depth, and resolution using a set of predefined scaling coefficients.

(3) ConvNeXt Liu et al. (2022b), 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 4.1, 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 datasets from 5.4, 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.

4.1.1 Ensemble Model Selection based on Mean Correlation Matrix

The goal of our strategy inspired in Nikita Kozodoi (2020) 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

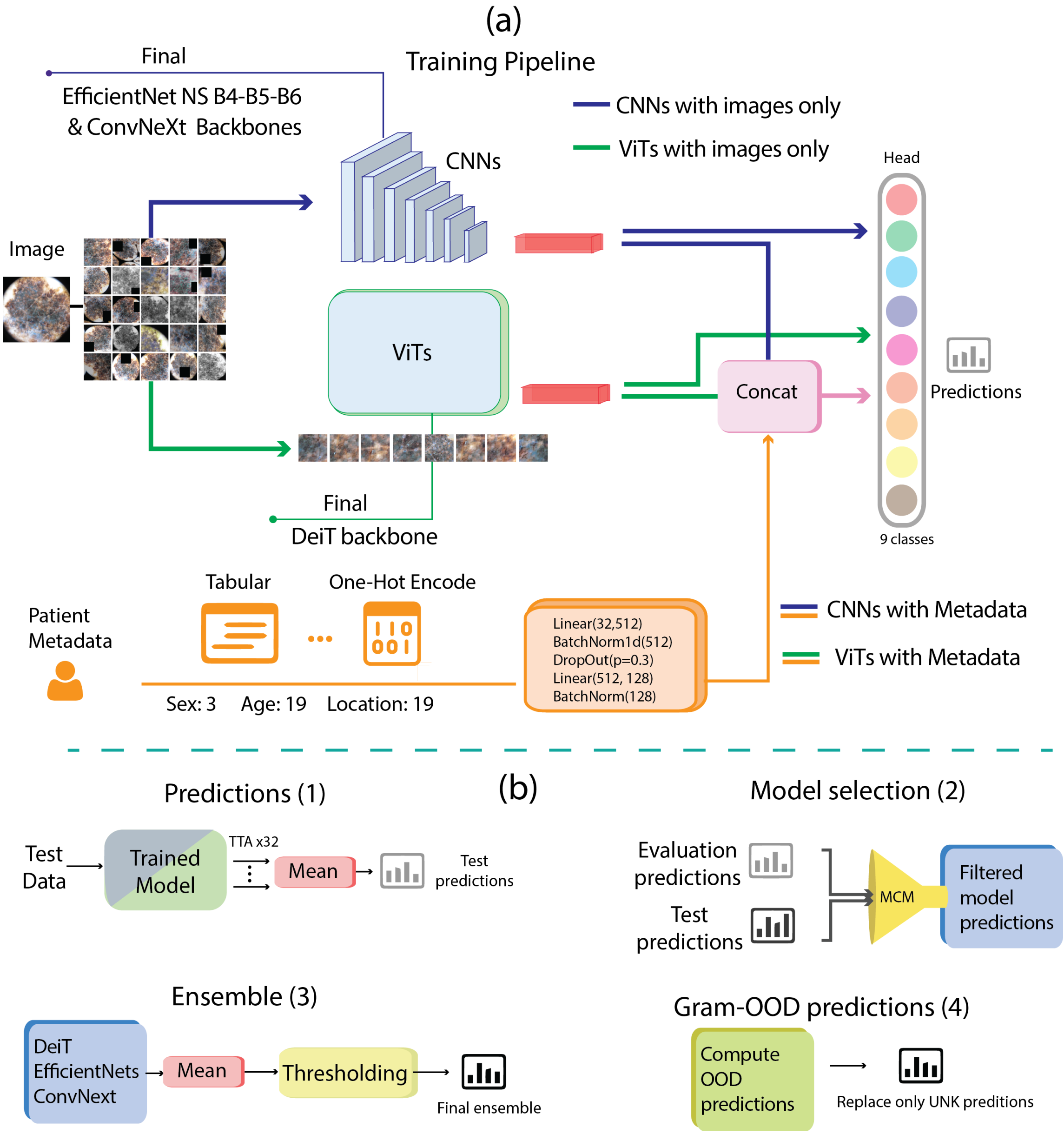


Figure 4.1: Diagram of the pipeline of our model. (a) depicts the training pipeline and (b) testing pipeline. The final ensemble is trained on the datasets addressed in 5, and had used both external images only, and metadata for both networks. The testing pipeline shows the generation of predictions in four stages

the difference in the arithmetic mean on each class correlation. Equation (4.1) 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 (4.2) 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.

$$\rho(x, y)^C_{v|t} = \frac{\sum[(x_i - y)(y_i - y)]}{\sigma x * \sigma y} \tag{4.1}$$

$$MCM = \frac{1}{9}\sum_{C=1}^{9} G_C, \qquad G_C = \left|\rho^C_v - \rho^C_t\right| \tag{4.2}$$

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 6.9 shows the application of the MCM in the model selection.

4.1.2 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* from Section 3.5, it was discovered that computing feature maps from convolutional layers rather than activation functions, see Table 4.1, could result in a slight improvement while retaining the pairwise correlations and layerwise deviation computation from the original method Sastry and Oore (2019) and the normalization extension proposed in Pacheco et al. (2020).

Method	TNR	AUROC	DTACC	AUIN	AUOUT
Gram-OOD* Pacheco et al. (2020)	7.028	45.456	51.311	18.628	78.163
Gram-OOD (Ours)	**9.226**	**59.414**	**57.083**	**26.793**	**83.205**

Table 4.1: Comparison of the usage of convolutional layers vs the activation functions as feature maps

4.1.3 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) Aurelio et al. (2019), and Focal loss (FL) Lin et al. (2020), were addressed in order to find the best pipeline, see Table 6.7. Equation 4.3 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.

$$l(x, y) = L = l_1, ..., l_N{}^\top, \qquad l_n = -\sum_{c=1}^{C} w_c log \frac{exp(x_{nc})}{\sum_{c=1}^{C} exp(x_{n,i})} y_{n,c} \tag{4.3}$$

Equation 4.4, 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.

$$FL = -\sum_{i=1}^{n} \alpha_i (i - p_i)^\gamma log_b(p_i) \tag{4.4}$$

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 Buda et al. (2018). This approach applied in Steppan and Hanke (2021) has demonstrated to significantly increase the performance in imbalanced datasets by a class probability distribution approximation. Richard and Lippmann (1991) 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 (4.5), 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:

$$p(c|x) = \frac{p(c)p(c|x)}{p(x)}, \qquad p(c) = \frac{|c|}{\sum_k |k|} \tag{4.5}$$

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. The factor can be seen in Figure 4.2.

Class	**AK**	**BCC**	**BKL**	**DF**	**MEL**	**NV**	**SCC**	**UNK**	**VASC**
Re-scaling factor	37.77	9.69	11.98	111.38	6.67	2.38	52.14	5.49	116.12

Table 4.2: Re-escaling factor given the probability distribution of the data.

4.2 Model Implementation

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

4.2.1 Optimizer and Learning Rate

After the comparison of state-of-the-art optimizers in Heo et al. (2021), 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 Morris (2018).

However, before making a choice on the LR, the configuration had to be selected based on the LR scheduler from section 4.2.2. In Smith (2018) 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.

4.2.2 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 Smith (2018) requires defining a minimum and maximum LR, to achieve the super-convergence. One cycle is shown in Figure 4.2, 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 6.4 shows the other LR schedulers tested.

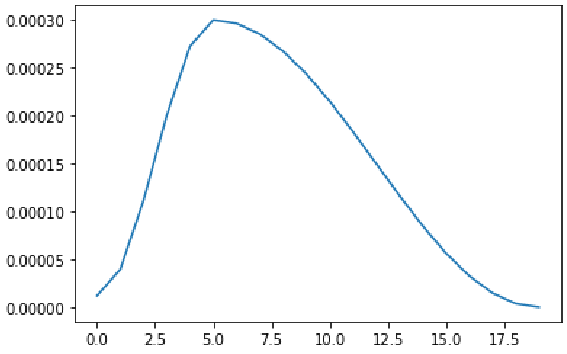


Figure 4.2: One Cycle LR.

4.2.3 Data Preparation

The images in the dataset are all from different sources, scanned at various resolutions and on the same color space. However, some of them are composed of microscope-like image cropping that were detected as outliers in section 5.2.4, and were preprocessed to see whether they could result in an generalized improvement as Gessert et al. (2020) 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 4.3 presents a few examples of all the 9577 images determined as outliers.

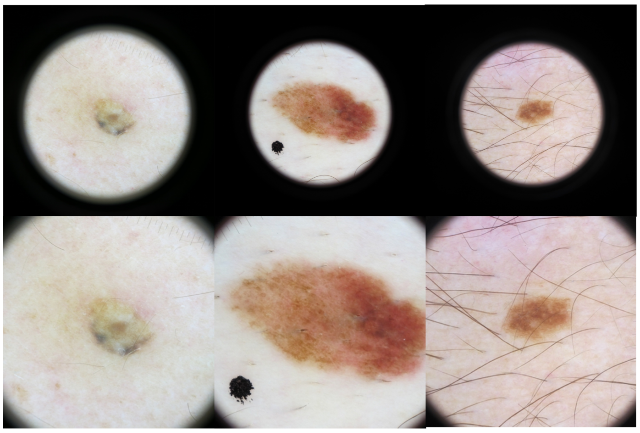


Figure 4.3: Preprocesssing of outlier images.

Additionally, it was essential to remove the missing values that were discovered in 5.2.1 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 4.3 depicts the amount of parameters for the tabular data.

Metadata	Number of parameters
Sex	3
Age	19
Anatomical Location	10

Table 4.3: Metadata number of parameters used as input for the models

4.2.4 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 Cubuk et al. (2019), RandAugment Cubuk et al. (2020) and AugMix Hendrycks et al. (2020). Before a selection, an adaptation of the customized standard augmentation by Ha et al. (2020) was compared in the section 6, 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.

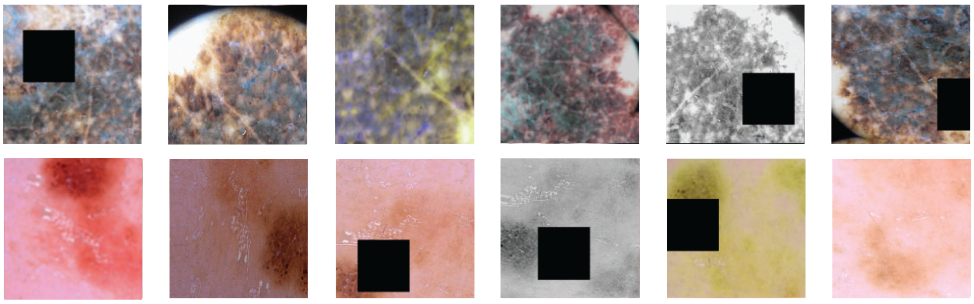


Figure 4.4: Image augmentation employed: a standard augmentation regime (random flip, rotation, brightness/contrast and blur/gaussian noise) followed by a random and resized crop strategy, CutOut of 30% image size, and gray and color-jitter/hue-saturation changes. Details can be found in 6.6.

Figure 4.4 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 Ali et al. (2020): 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 inspiration from Contrastive Learning Chen et al. (2020) 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 Devries and Taylor (2017) 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 Gessert et al. (2020), were explored, but yielded no benefits and were therefore rejected. Appendix 6.6 has the whole augmentation configuration, and Appendix A.1 contains image samples of the augmentation tested.

Chapter 5

Datasets

5.1 Melanoma Detection and Characterization with Deep Learning

The conventional and primarily method of diagnosing skin cancer is by visual inspection, which may be supplemented by dermoscopic analysis, a biopsy, and histopatological evaluation Esteva et al. (2017). Essentially, the asymmetry, boundary, severity, and physical size of a skin lesion define its type Sharmeela and Asha (2017). However, there is a restriction that bears the highest weight and deserves to be evaluated; melanoma is distinct from other skin cancers in that it is usually a proliferation of pigmented cells, although not in all cases, which makes it difficult to address as a unitary and easy diagnosis. It can also originate in different sites of the body, including the head and neck, as well as the bottom of the palms and souls (see Table 5.2), which usually does not occur in other types of skin cancer and likewise deserves attention.

All of these criteria must be considered in order to develop a powerful pipeline capable of performing both on the multi-class classification task of skin lesion from the 2019 ISIC challenge and the singular malignant melanoma prediction from the 2020 dataset. Although melanoma is not the most common type of cancer, it is the deadliest one since it is the skin cancer that is most prone to spread to other organs of the body CDC. (2022). According to the research presented in Foundation (2022), it is more common in older people, although it is far more common in younger than other types of fatal tumors, and yet it is more common in men overall, but rates are higher in women before age 50. All prior assertions allow for potential information to enrich a DL model; patient-level information such as age, sex, and anatomical location.

5.2 Data Properties and Exploratory Data Analysis

5.2.1 Missing Values

As it is frequent in any tabular environment, missing values are a common problem which arises from typing errors and circumstances of the natural world. However, tolerances to this pitfall should be raised in advance and the missing values should be properly identified and handled to produce a more robust result. Figure 5.1 shows a data-dense nullity matrix to visually see where the null values are found in the dataset.

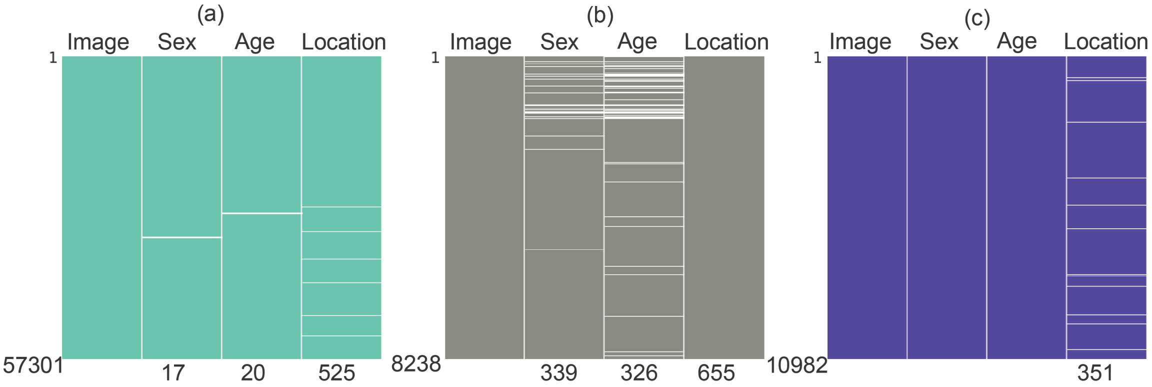


Figure 5.1: Missing values from the combined datasets. (a) Display of the whole training data from 2019 and 2020 datasets, while (b) from 2019 test data, and (c) 2020 test data. The white lines represent missing data.

5.2.2 Data Distribution

The examination of the distribution of the target diagnosis, and the location can be seen in Figure 5.2. Once again, as in Tables 5.1 and 5.2, the imbalance in the class label is evidently seen in the distribution plot (a), which showcases the unknown as the majority class (46%), followed by the NV with the second largest proportion (31%) from the totality of 57,301. MEL appears as the third class with a significantly lower number of samples 5,049 which accounts for about 9%. Similarly in proportion, the BKL and BCC appear with 6% and 5% accordingly, while the minority classes SCC, VASC and DF with about 1%, 0.6% and 0.4% respectively. Note that in Table 5.1, the abbreviation of the classes are shown.

Following Figure 5.2, the anatomical location in (b) also portraits an imbalance scenario, with the torso and lower extremity filling more than half the dataset; 29%

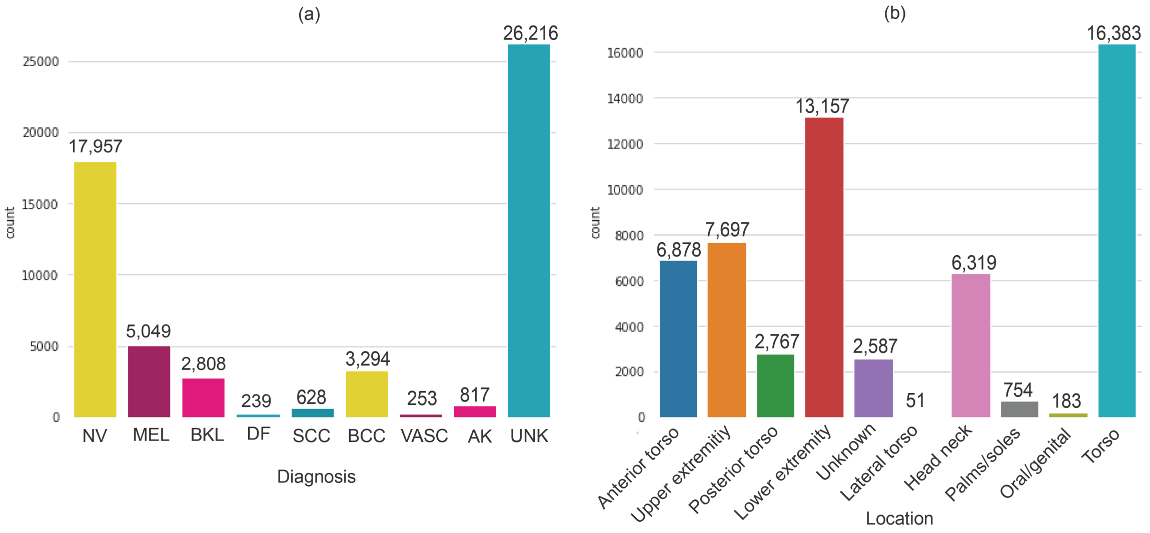


Figure 5.2: Distribution table: (a) shows the diagnosis distribution in the whole dataset, (b) portraits anatomical location count of the whole dataset.

and 23%. The upper extremity, anterior torso and head/neck show the second largest proportion with 13%, 12% and 11%, respectively. The posterior torso has a similar proportion with the number of unknown entries of about 5%. Finally, the palms/soles, oral/genital and lateral torso hold the minority class proportion with 1% or less.

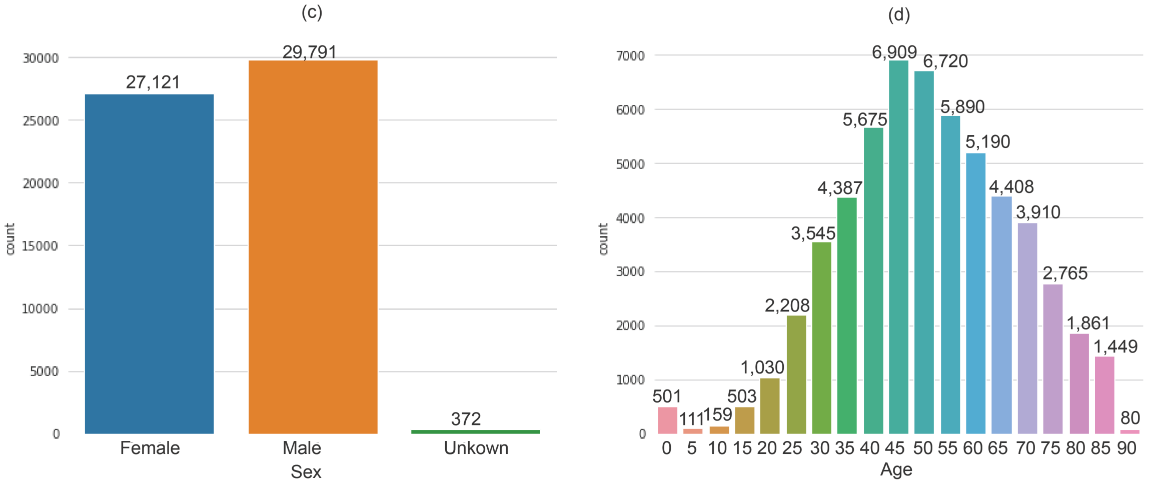


Figure 5.3: Sex and age distribution: (c) shows the sex distribution from the whole dataset, and (d) the age distribution count of the whole dataset.

Similarly, Figure 5.3 represents the patients' age and sex distributions. The biological sex count is presented in graph (c), with a modest differential of just around

5% greater for the male sex. Luckily, the unknown samples constitute fewer than 1% of the total samples, because in Leonardi et al. (2018), it is stated that when analyzing incidence data in relation to sex, women are more frequent in younger aged groups, while the male sex prevails from the age of 55 onwards. In figure 5.3, (d) shows the entirety of the age distribution count, with the reduction of skin lesion cases in older ages and an increase for the younger population between 25 and 40 years old. Again, Leonardi et al. has indicated that incidence grows linearly after the age of 25 until the age of 50, and subsequently declines, particularly in the female sex. Similarly, metadata reveals that the mid-age group has the largest frequency, peaking at 45 years old and declining until the elderly, implying that new information can be derived from it.

5.2.3 Data Relations

A first relevant subject for analysis is the relation between the anatomical location and sex, since skin cancer is mainly driven by UV radiation from the sun, and generally men and women have a different dressing habits, which could result in different exposure rates and rich features for a DL model. Figure 5.4, draws the proportion of sex in each anatomical location. The proportion of skin cancer remains broadly comparable, with the exception of a slightly, but significant proportion for the head/neck and torso —both anterior and posterior— in the male sex, accounted for 57%, 55%, 57%, and 60%, and a greater proportion for the upper and lower extremities in women, with 52% and 55%, respectively.

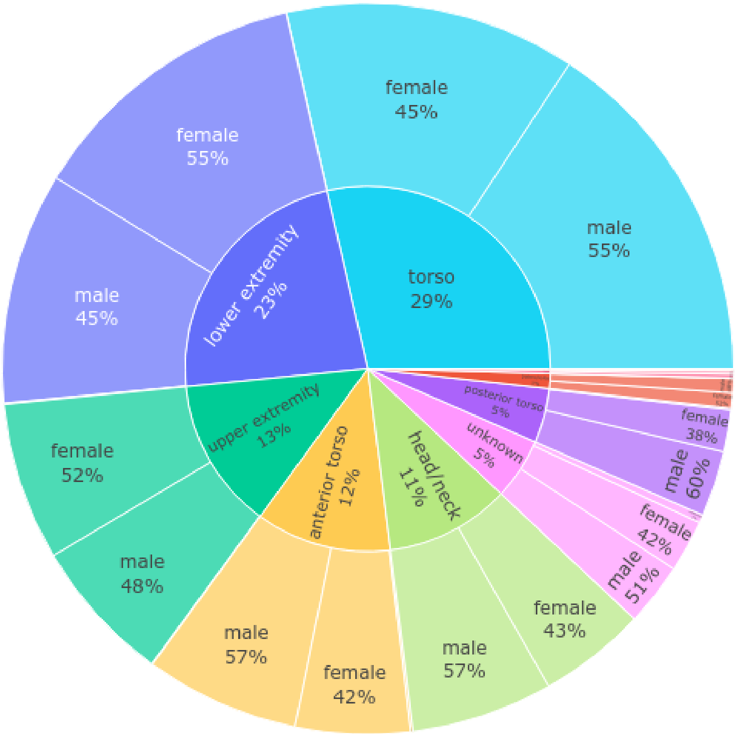


Figure 5.4: Stratified hierarchical visualization of the sex proportion in the anatomical location.

One thing to bear in mind is if there is a prevalence of any particular diagnosis depending on the age of the patient. The relation between the age vs. each class can be seen in the density Figure 5.5. Interestingly, it exhibits two primary density patterns; one of which indicates a preponderance of NV, DF, and UNK from the age of 20 to a first peak around 45, after which it drops over the next years until old age. An initial peak on the MEL is depicted. However, the most of the occurrences do not arise here. The other density cummulus is rising from the middle-age to the elderly population, along with remaining classes: SCC, AK, BKL, BCC and VASC. It is to note that the MEL also shows its highest peak in older population.

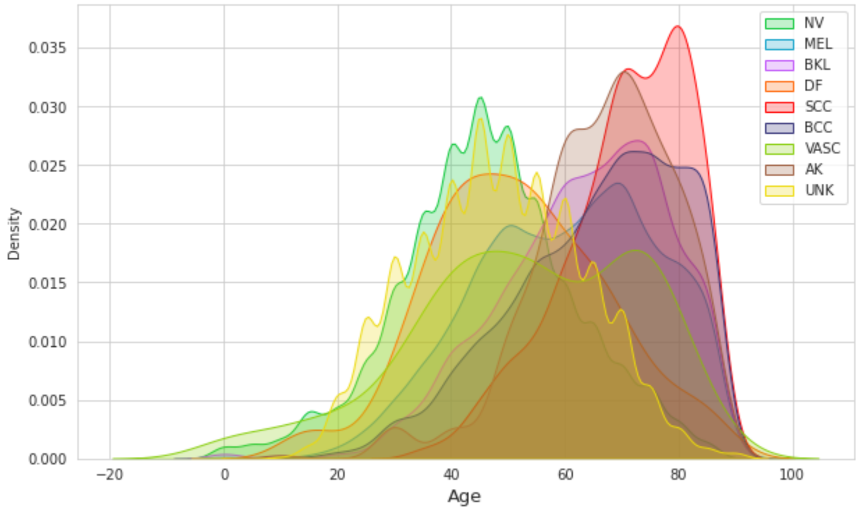


Figure 5.5: Age distribution per class

To dive further into the MEL, Figure 5.6 outlines notably the prevalence of melanoma in the male sex; both in the anterior torso and head/neck than in the upper/lower extremities. In women, the lower extremity has the highest prevalence of melanoma, followed by the anterior torso and with the upper extremity 5% higher.

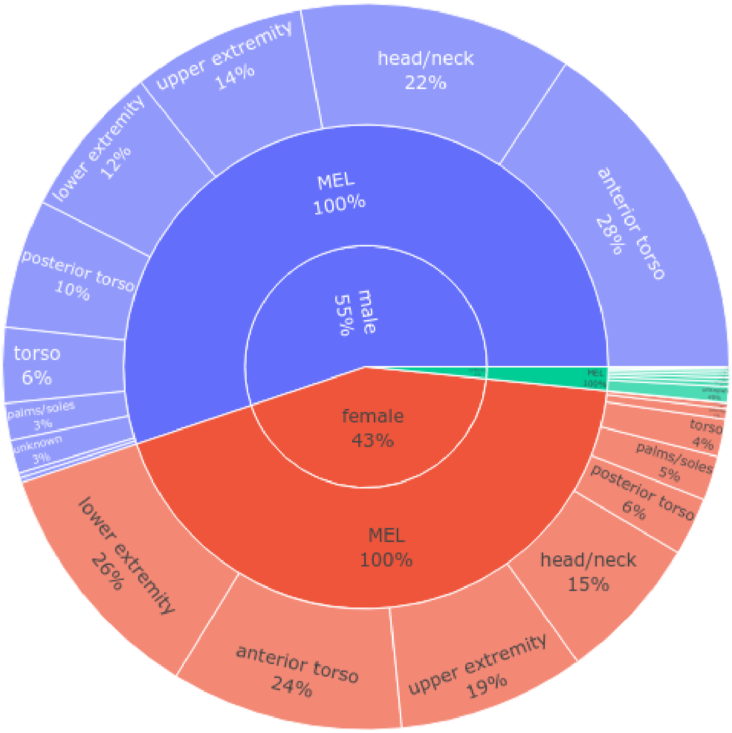


Figure 5.6: Melanoma relation between the anatomical location for both sexes.

Overall, the metadata seems to provide new features and suggests that it would be worth training a classifier using both the metadata and the skin disease image to assess how well that combination could perform.

5.2.4 Outlier Detection

The presence of outliers is a significant feature that needs to be addressed in any dataset, especially if OOD is of interest Yang et al. (2021). One typical method for identifying outliers, which is suitable in this setting, is to observe the mean intensity level and the standard deviation in order to later verify their boundaries for outliers on the body of the dataset. Figure 5.7 shows a 2-dimensional setting with data points indicating the mean and standard deviation for each image.

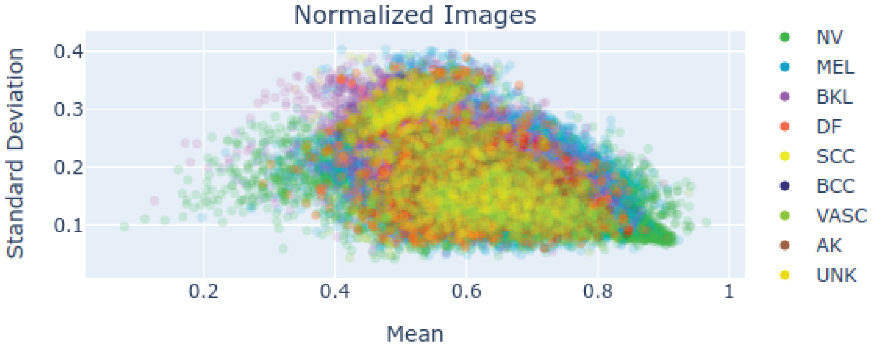


Figure 5.7: Outliers search with the mean and standard deviation of the normalized images.

Regardless of the density aggravated by the proportion of the majority classes, there are notable distinctions in the mean intensity values of the different data points, and particularly there are a number of outliers visible in the domain outside of the 0.4 for the mean, which can be scrutinized further using the box plot per class depicted in Figure 5.8. Moreover, this illustration depicts numerically distant observations from the rest of the data. Indeed, for NV and UNK, a large number of data points are shown outside the whiskers of the box plot for each class, allowing thresholds to be discovered for outlier identification. Furthermore, Figure 5.9 (a) shows the distribution of outliers present for all nine classes; a total of 9577 after identification, and (b) renders visually some of them.

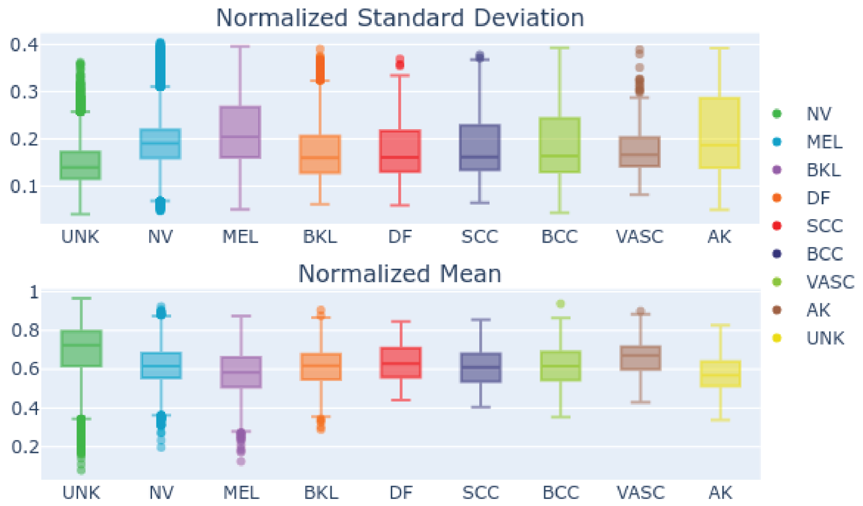


Figure 5.8: Outliers box plot with the normalized mean and standard deviation for each class.

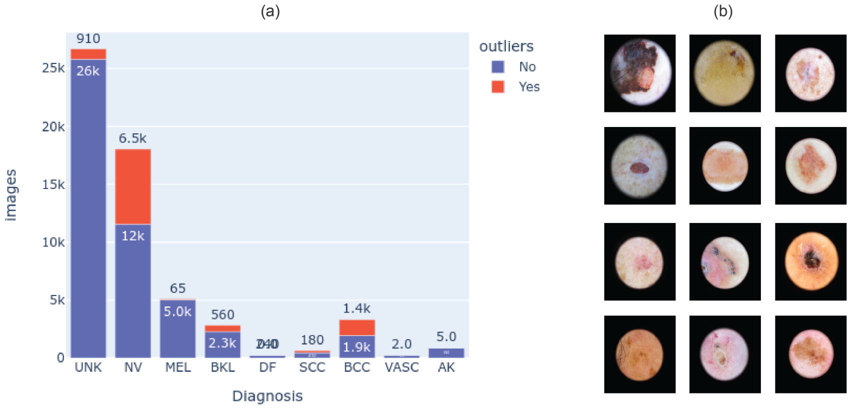


Figure 5.9: (a) Outliers distribution per class; (b) a few examples of the outliers found.

5.2.5 t-SNE

A low-dimensional representation of high-dimensional data is created using the technique known as t-SNE for visualizing feature vectors. Here, colored point clouds that represent the various disease categories show how the t-SNE algorithm groups the skin lesions clusters. Figure 5.10 holds the representation made after several executions varying the algorithm hyper-parameters. It should be noted that both the class imbalance and the natural resemblance of the skin lesions (particularly with the unknown label) significantly contribute to the complexity of this real classification problem.

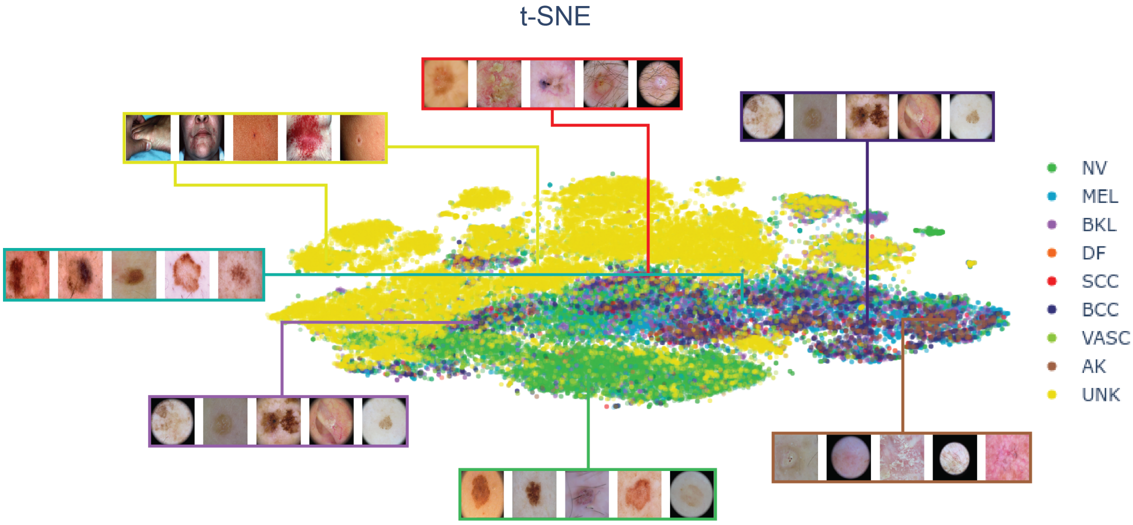


Figure 5.10: t-SNE visualization of the whole dataset.

5.3 The International Skin Imaging Collaboration (ISIC) project

The data utilized in this study primarily comprises both the ISIC 2019 dataset Tschandl et al. (2018) Gutman et al. (2018) Combalia et al. (2019) and the ISIC 2020 dataset Rotemberg et al. (2021). The ISIC datasets are a compendium of skin lesion images that have been annotated with a number of labels, including metadata. During live competitions that went until August 23 of 2019, and ran from May 27 of 2020, to August 20 of 2020, both datasets respectively were made available to the general public for download via the Kaggle platform. Note that 2019-2020 datasets are distributed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) via the ISIC Archive.

Given the absence of dermatologic imaging standards, the effectiveness and quality of skin lesion imaging are currently compromised. ISIC therefore, focuses on creating suggested standards to improve the digital skin images' quality, privacy, and share-ability producing tools for the computer science and dermatology communities, such as an open source public access repository of skin images (ISIC archive). This repository provides free access to photos that may be used in research, education, and the creation and sponsoring of Grand Challenges in artificial intelligence for the evaluation of diagnostic artificial intelligence systems.

5.3.1 The ISIC Archive

Over 150,000 photos in total, of which 70,000 have been made publicly available via the open source ISIC Archive platform. Metadata describing extra attributes at the image level are linked to images; Malignant versus benign (which also includes values ″indeterminate″ and ″unknown″) diagnosis, estimated age, sex, anatomic location (where the lesion is located on the body with 7 sub-categories) clinical size, diagnostic type melanoma related material, such as personal and family histories and other information (class, mitotic index, thickness, ulceration) is related among images. It should be noted that not all of the aforementioned metadata is available in all datasets, and missing values are to be expected.

5.3.2 The ISIC Challenges

ISIC has sponsored annual computer vision challenges since 2016, using high-quality human-validated training and test sets of thousands of CC-0-licensed images and metadata. In 2018, the leading algorithms were consistently outperforming clinicians in ″reader studies″. Additional challenges in 2019 and 2020 were designed to address the OOD problem, and asses the impact of clinical context, allowing researchers and students continual benchmarking of the algorithm's performance.

5.3.3 The ISIC 2019 dataset

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 Combalia et al. (2019) and the HAM10000 dataset from the Department of Dermatology, Medical University of Vienna Tschandl et al. (2018). Table 5.1 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. Figure A.6. depicts a few skin disease samples from the ISIC 2019 dataset.

Class diagnosis	Abbreviation	Samples
Melanocytic nevus	NV	12875 (50%)
Melanoma	MEL	4522 (18%)
Basal cell carcinoma	BCC	3323 (13%)
Benign keratosis	BKL	2624 (10%)
Actinic keratosis	AK	867 (4%)
Squamous cell carcinoma	SCC	628 (3%)
Vascular lesion	VASC	253 (1%)
Dermatofibroma	DF	239 (1%)
Outlier class	UNK	0 (0%)
Total		25331 (100%)

Table 5.1: Diagnosis distribution for 2019 dataset.

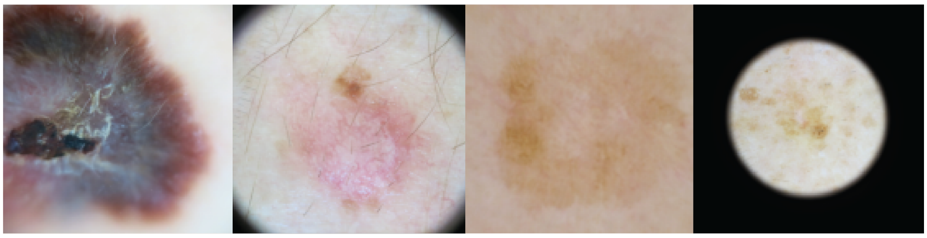


Figure 5.11: Sample skin lesions in ISIC 2019 dataset.

5.3.4 The ISIC 2020 dataset

ISIC 2020 dataset Rotemberg et al. (2021) 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 Rotemberg et al. (2021). In contrast to the 2019 dataset (5.3.3), the unknown 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. Table 5.2 shows the four diagnosis labeled in the dataset with the samples.

The aim for this challenge was to predict a binary target for each image; the only thing that matters is whether the skin lesion is benign or malignant, that is, melanoma or not. It is presented as the first melanoma and comparative lesions dataset from the same patient and boosts new ML challenges, particularly in Melanoma research. The same metadata than in the previous year was available

Class diagnosis	Abbreviation	Samples
Melanocytic nevus	NV	5193 (15%)
Melanoma	MEL	584 (2%)
Benign keratosis	BKL	223 (1%)
Cafe-au-lait macule atypical melanocytic proliferation	UNK	27126 (82%)
Total		33126 (100%)

Table 5.2: Diagnosis distribution for the 2020 ISIC dataset.

for this dataset. Figure 5.12 shows a few skin lesion examples.

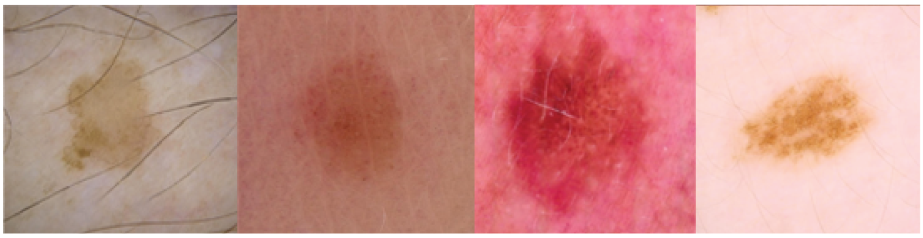


Figure 5.12: Sample skin lesions in ISIC 2020 dataset

5.3.5 Duplicates

The competitor host on an official Kaggle article Weber (2020) confirms there are true duplicates in the 2020 ISIC dataset. Because the dataset is comprised of a few samples of 2019 ISIC data as well as the 2020 dataset, detecting duplicates is critical to avoiding leaks on validation data and thus, is the first step to take when attempting to develop a solid generalization. Nonetheless, finding duplicates in a dataset is an arduous task, and there are several ways available, ranging from RAPIDS cuML kNN with image embeddings Deotte (2020), to the most basic approaches such as scanning for a file size, resolution, or file name. A simple and effective method is to employ hash functions, which transform or translate the input data to a fixed-length string that may be thought of as the ”fingerprint” or ”signature” of the input data. Finally, in the entire dataset, up to 1162 duplicates were found using the perceptual hash method Zauner (2010), and as a result, the validation process is expected to be greatly enhanced by avoiding duplication. Figure 5.13 shows a few duplicates instances discovered.

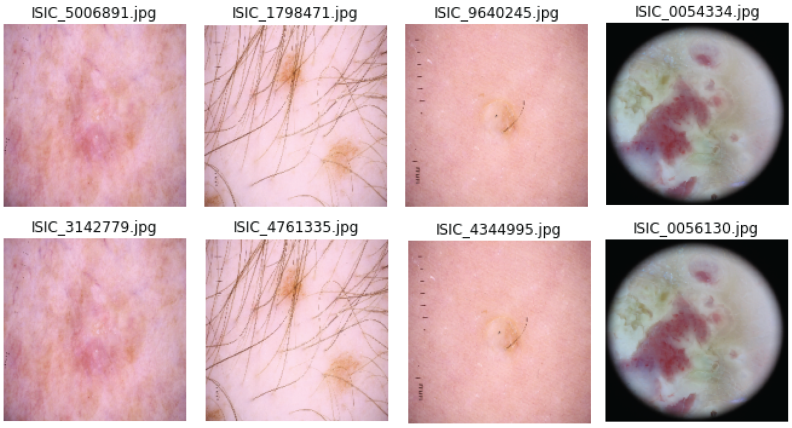


Figure 5.13: Examples of duplicate samples in the 2019 and 2020 ISIC datasets found with the PHash algorithm.

5.3.6 Patient-Level Metadata

The same patient-level contextual information has been made accessible for researchers to use in their models for the 2019 and 2020 ISIC competitions, with special prizes awarded to the best scoring models as an incentive for its use in both cases. It is worth mentioning that not all images contain meta-data, as it came to light in section 5.3.1 and 5.2. The metadata available comprises the patient's biological sex, age, and the location of the skin disease on the body. The main difference in this regard between the datasets is that the 2020 dataset includes an extra component associating the skin lesion with the patient-id; with over 2000 patients in total for 2020 ISIC dataset. The metadata shared and used for both tasks is depicted in details in Table 5.3.

Metadata	Detalis
Age	5 to 85 range Increments of 5
Sex	Female Male
Skin Anatomical Location	Torso Lower extremity Upper extremity Anterior torso Head/neck Posterior torso Palms/soles Oral/genital Lateral torso

Table 5.3: Patient-level information with details, both from the 2019 and 2020 ISIC datasets.

5.4 External Data

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 Steppan and Hanke (2021), the outlier class for training was addressed through the usage of a subset of a collection of datasets, which are detailed below:

5.4.1 SD-198

SD-198 is a large and widely available clinical skin disease dataset that has been collected and made public. In contrast to dermoscopic images, clinical images are captured by dermatologists using standard camera equipment such as smartphones or digital cameras and under varying lighting conditions that are not consistently sustained Rotemberg et al. Sun et al. (2016). The dataset contains 6,584 clinical images covering 198 distinct skin disorders with associated metadata, which vary according to scale, color, shape and structure. A glimpse of the images used can be seen in Figure 5.14.

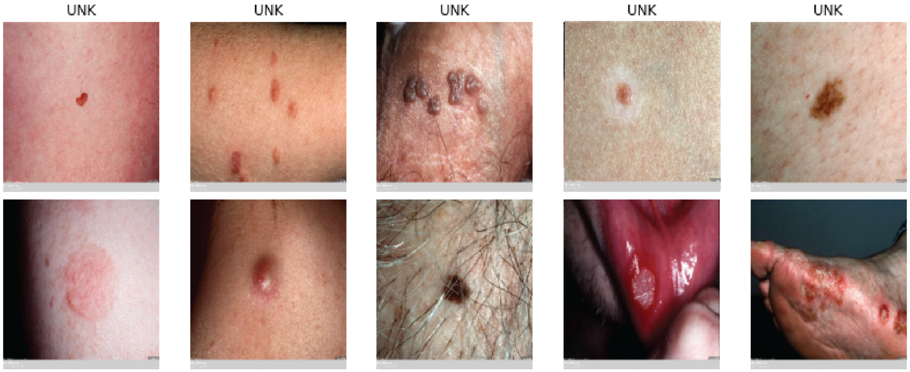


Figure 5.14: Sample of clinical skin lesions from the SD-198 dataset.

As in Steppan and Hanke (2021), it is used as a foundation for the unknown class and particularly, this dataset was employed as the majority source for the outlier class from the 2019 ISIC Challenge, taking 5,944 in total, whose ground-truth class is unknown.

5.4.2 PH2 Database

The study given in Mendonça et al. (2013) provides the PH2 database of 200 dermoscopic images, which is freely available and was obtained at Pedro Hispano

Hospital. This database contains medical annotations for all images, such as lesion segmentation, clinical diagnosis, and dermoscopic criteria (asymmetry, colors and the presence of typical and atypical differential structures). These images comprise 80 common nevi, 80 atypical nevi, and 40 melanomas each of a resolution of 768x560 pixels. For external data contribution, a total of 160 nevus (NV) and 40 melanoma (MEL) were included in the external data. Figure 5.15 shows an example of the images utilized.

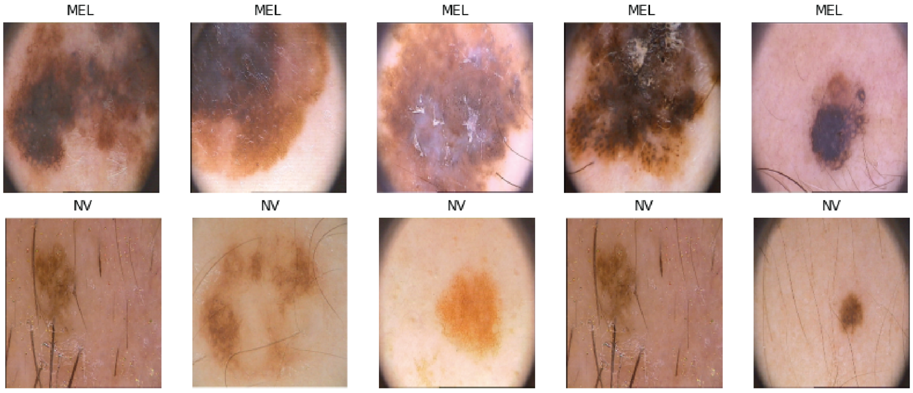


Figure 5.15: Sample of clinical skin lesions from PH2 database.

5.4.3 7-Point Criteria Evaluation Database

With the study made in Kawahara et al. (2019), a database is utilized to evaluate the computerized image-based prediction of the 7-point checklist for malignant skin lesions. The 7-point checklist (7PCL) was created in Glasgow in the 1980s to enable non-dermatologists recognize characteristics suggesting melanoma Walter et al. (2013). The collection contains roughly 2000 clinical and dermoscopic RGB images, as well as structured metadata suitable for training and testing computer assisted diagnostic (CAD) systems. Figure 5.16 shows a selection of the images used.

The 7-point dataset contributed to a total of seven classes, nevus 569 (NV), 252 melanoma (MEL), 85 benign keratosis (BKL), 42 basal cell carcinoma (BCC), 29 vascular lesion (VASC), 20 dermatofibroma (DF), and finally 14 outlier classes (UNK).

5.4.4 Light Field Image Dataset of Skin Lesions

The work from Rerábek and Ebrahimi (2016) presents a contribution for the research community, in the form of a publicly available light field image datasets of

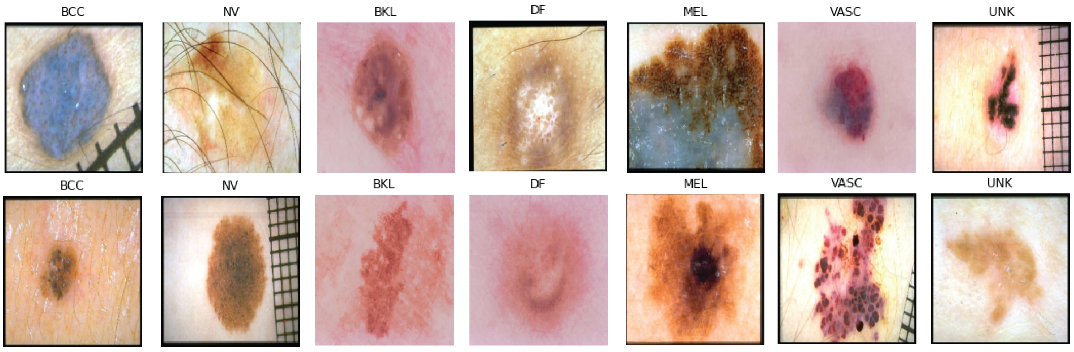


Figure 5.16: Sample of clinical skin lesions from the 7-point criteria database.

skin lesions named SKINL2. The datasets contains light fields images, captured with a focused plenoptic camera and classified into eight clinical categories, according to the type of lesion. However, italso includes the dermatoscopic image of each lesion which were then ones taken from this database. This dataset has as made with the motivation for advancing medical imaging research and development of new classification algorithms based on light fields and dermatology studies.

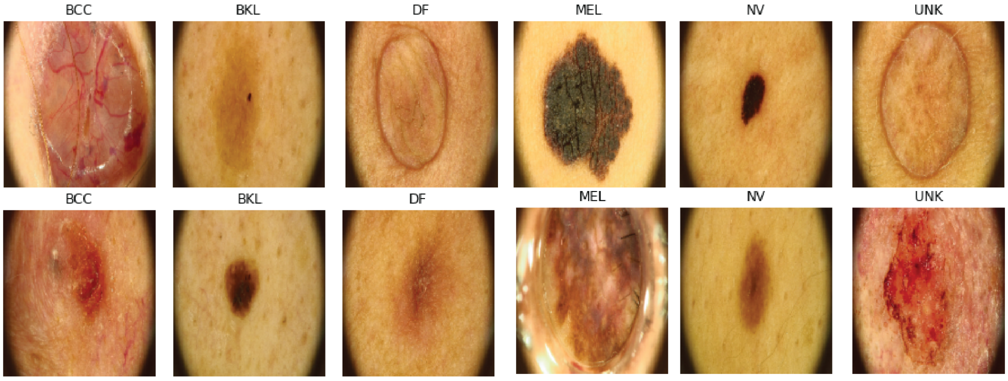


Figure 5.17: SKINL2 v1-2-3 dataset samples

The dataset SKINL2 has three different versions that were added in order to improve the work in Kawahara et al. (2019). The dataset contributed to a total of seven classes, particularly regarding 128 nevus (NV), 53 benign keratosis (BKL), 39 basal cell carcinoma (BCC), 30 melanoma (MEL), 17 dermatofibroma (DF), and finally 32 outlier classes (UNK). Figure 5.17 shows a selection of the images used.

5.4.5 MED-NODE

Non-dermoscopic digital images of lesions were contributed in Giotis et al. (2015). This dataset contains 70 melanoma and 100 nevus images from the digital image archive of the Department of Dermatology of the University Medical Center Groningen (UMCG), which were used for the development and testing of the MED-NODE system for skin cancer detection. It automatically extracts lesion regions and then computes descriptors regarding color and texture. The external data now contains all 170 photos. Figure 5.18 illustrates some of the photos that were utilized.

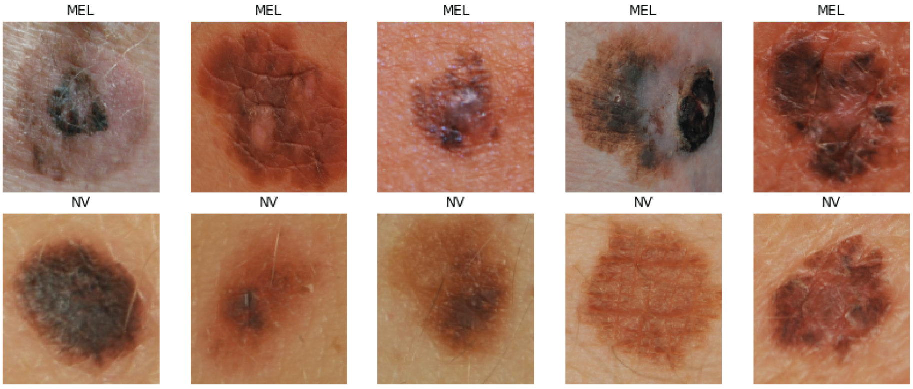


Figure 5.18: Sample of clinical skin lesions from MED-NODE criteria database.

Chapter 6

Validation

In this chapter, 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.

6.1 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 6.1, 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 Archive (2019):

- 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 6.1: Metrics defined by the 2019 ISIC live challenge to assess models performance.

6.2 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 Paszke et al. (2019), 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 from section 3.6 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 Wightman (2019), 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 Falcon (2022), and Wandb Biewald (2022), 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

6.3 Baseline and Default Settings

In order to get a decent start, the results from the CNNs baseline in research Steppan and Hanke (2021) were adopted and with the configurations made from the training and computational limitations mentioned in Section 6.2. 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 Tan and Le (2019), Inception Resnet V2 Szegedy et al. (2017) and ResNeXt Xie et al. (2017). In the case of ViTs used as baseline: the basic ViT Dosovitskiy et al. (2021), BEiT Bao et al. (2022), SwinT Liu et al. (2021), and SwinTV2 Liu et al. (2022a). Hence, the relevant models and their performance are displayed in Table 6.2. Furthermore, initially only images from the whole dataset shown in Figure 6.3 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.

Method	# Params	Image size	Data usage	Val BACC	2019 Score
SWSL ResNeXt-101 32x4d Yalniz et al. (2019)	54M	224	External	72.09%	0.429
Inception-ResNet-V2 Szegedy et al. (2017)	56M	299	External	76.33%	0.433
EfficientNet b4 Tan and Le (2019)	19M	380	External	71.11%	0.424
EfficientNet b5 Tan and Le (2019)	30M	456	External	77.73%	**0.483**
CNNs baseline ensemble	**0.496**				
ViT-L-16 Dosovitskiy et al. (2021)	304M	224	External	75.73%	0.418
Swin-L-4 Liu et al. (2021)	197M	224	External	73.02%	**0.464**
SwinV2-B- Liu et al. (2022a)	88M	256	External	74.56%	0.412
BeiT-B-16 Bao et al. (2022)	87M	224	External	75.13%	0.403
ViTs baseline ensemble	0.482				

Table 6.2: ISIC 2019 score and BACC baseline. Note that there is no data preprocessing, duplicates removal or imbalance handling.

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 6.2 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 Section 5.2.4, 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.

6.3.1 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 4.2.2, and AdamP 4.2.1, 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 4.2.4. The final configuration of our ensemble is given in Table 6.3.

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	[6.6]

Table 6.3: Training and hyper-parameter configuration for the final models

6.4 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 discussed in 3.6, with four optimizers; Cosine Annealing and SGD Cosine Annealing with warm restart Loshchilov and Hutter (2017), SGD with Cyclic LR Smith (2017), and straightforward LR step decay with AdamP. Figure 6.1, shows a diagram of their behavior adjusting the LR during training. The strategies assesst in the experiments can be found in Table 6.4. 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.

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

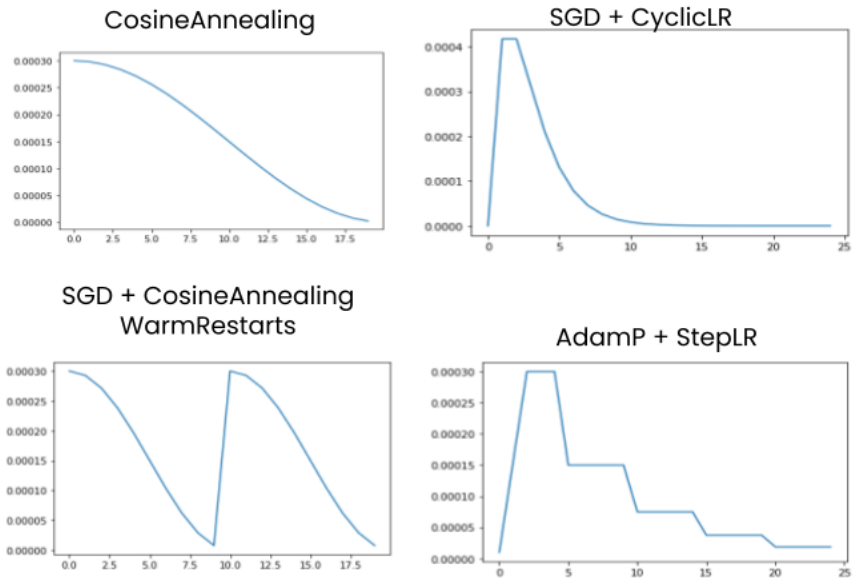


Figure 6.1: Diagram of popular optimizer and LR scheduler pairs behavior adjusting the LR during each epoch of training. All of them followng the principle of super-convergence

On the other hand, the more recent Sharpness-Aware Minimization (SAM) optimizer Foret et al. (2021) suggested it promising results improving generalization without strong data augmentation Chen et al. (2022). However, the experiment presented in Figure 6.2 corroborated that SAM requires almost twice the training time to reach a comparable performance to AdamP optimizer, and since its usage is suggests with robust or semi-supervised methods, for this reason was thereby also discarded.

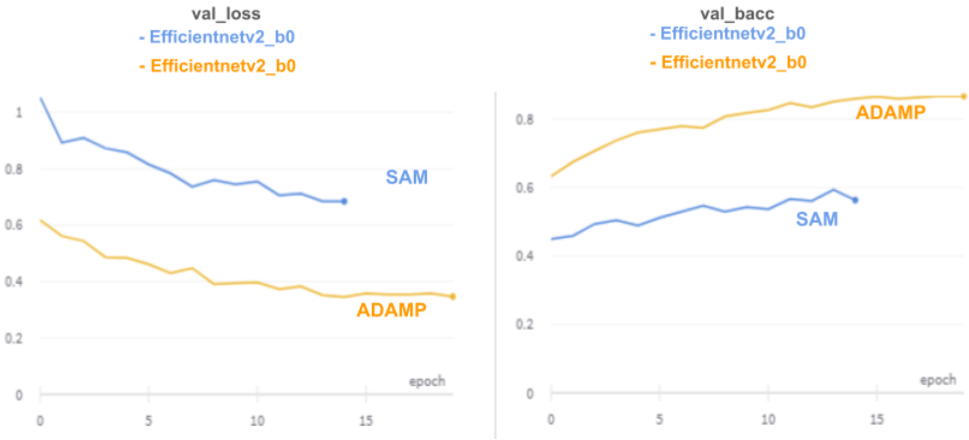


Figure 6.2: Run comparison of SAM and AdamP on the rapid and low parameter EfficientNetV2-B0. The tendency shows that SAM requires almost twice the time to reach comparable results to AdamP.

6.5 Data Splitting

For the data splitting the objective was find a strategy that could work for both model selection and hyper-parameter 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 Steppan and Hanke (2021), demonstrated a slight improvement for the outlier class. Therefore, datasets described in Figure 6.3 were used to feed the models in order to reach diversity in the DL ensemble. Moreover, in order to include metadata features from section 5.2, 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 6.4.

As a side note, if the goal is to improve the generalization 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 Anguita et al. (2012). Nonetheless, the fact that just by using the hold-out method our model could achieve top scores, demonstrates the potential and possibility of increasing

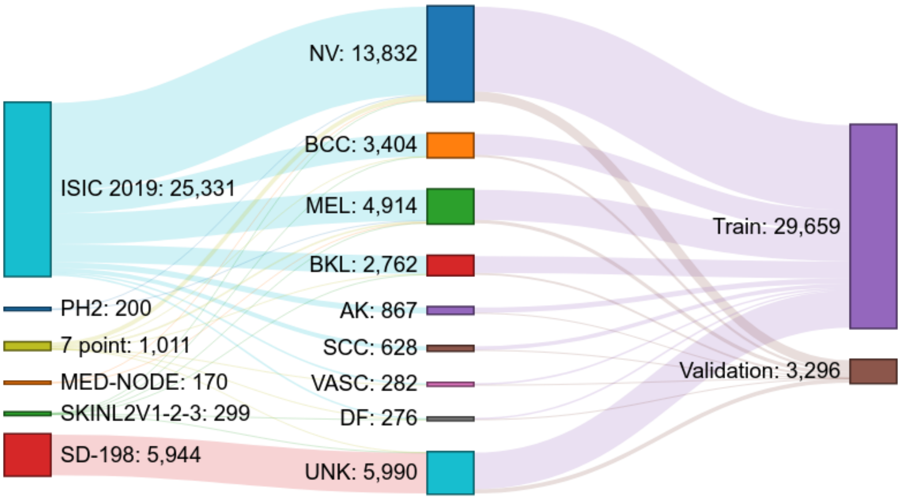


Figure 6.3: Skin Lesion Datasets Distribution for the external data. It displays the 25,331 samples from the ISIC 2019 as well as the contributions from the remaining external datasets and also indicates the splitting made for training and validation.

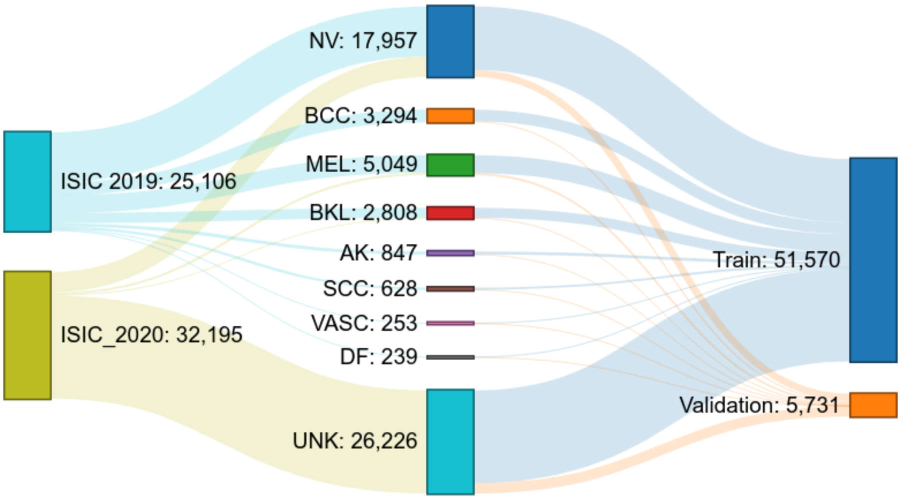


Figure 6.4: Metadata Skin Lesion Datasets Distribution for the 2019 and 2020 ISIC datasets. The contribution in each class is clearly demonstrated here, along with the splitting approach and proportions for training and validation.

the score further on the live leaderboard by simply integrating CV for model selection.

6.6 Data Augmentation Validation

We applied 13 techniques of data augmentation that are shown in Table 6.5. Table 6.6 compares the augmentation regime employed in this study to the alternative conventional augmentation methods. Different image sample representations may be found in Appendix A.1. The experiments used the same setup as in the baseline 6.3, 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.

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 6.5: Albumentation configuration for the training data.

Following the criteria from the melanoma ABCD rule Kasmi and Mokrani (2016), 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.

Method	**Model**	**Metric**	
		Val BACC	**2019 Score**
AugMix Hendrycks et al. (2020)	EfficientNetV2-B0 Swin-L-4	78.83% 76.14%	0.429 0.403
AutoAugment Cubuk et al. (2019)	EfficientNetV2-B0 Swin-L-4	79.35% 77.67%	0.434 0.552
RandAugment Cubuk et al. (2020)	EfficientNetV2-B0 Swin-L-4	80.07% 79.6%	0.439 0.419
Adapted Augmentation (Ours)	EfficientNetV2-B0 Swin-L-4	83.62% 83.87%	**0.495** **0.479**

Table 6.6: Data augmentation comparison results, using a ViT and a CNN for each data augmentation regime.

6.7 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 6.7 presents the comparative results of the three techniques indicated in 4.1.3 in order to analyze which approach among the conventional methods for handling imbalanced datasets in skin lesion classification should be preferred.

Imbalanced method	**Model**	**Metric**	
		Val BACC	**2019 Score**
Weighted Cross Entropy Aurelio et al. (2019)	EfficientNetV2-B0 Swin-L-4	84.34% 81.94%	0.511 0.504
Focal Loss Lin et al. (2020)	EfficientNetV2-B0 Swin-L-4	87.24% 86.75%	0.521 0.515
Thresholding (Ours)	EfficientNetV2-B0 Swin-L-4	86.94% 86.75%	**0.536** **0.526**

Table 6.7: Comparison of experimental results for Imbalanced methods

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.

6.8 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 6.8. Although BEiT-L is a powerful network for the ImageNet dataset, as demostrated by Bao et al. (2022), 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 6.9, depicts the ensemble methods chosen by verifying with three possibilities: (1) a rank of probabilities as used by 2.2.3.2, (2) majority voting scheme, and (3) model averaging, both discussed in 3.4; 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 Pacheco et al. (2020) and described in Section 4.1.2.

Table 6.10 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 6.5. It illustrates the new ROC Curve for the UNK class, alongside a dashed line corresponding to the previous ROC Curve (a) from Figure 6.9. The rest of the classes remain the same as the Gram-OOD only replace the predictions from the outlier unknown class.

Method	# Params	Image size	Data usage	Val BACC	2019 Score
ViT-L-16	26M	224	External	78.35%	0.514
Dosovitskiy et al. (2021)			Meta	83.56%	0.527
VOLO-D3	306M	512	External	82.31%	0.512
Yuan et al. (2022)			Meta	85.36%	0.516
DeiT-D3	305M	384	External	89.97%	0.592
Touvron et al. (2021a)			Meta	**91.73%**	**0.593**
CaiT-M-36	271M	380	External	84.29%	0.571
Touvron et al. (2021b)			Meta	88.21%	0.589
Swin-L-4	197M	224	External	81.17%	0.526
Liu et al. (2021)			Meta	83.87%	0.564
Swin-L-V2	197M	384	External	86.10%	0.563
Liu et al. (2022a)			Meta	89.46%	**0.610**
ViTs Ensemble	0.612				
SWSL ResNeXt-101 32x4d	54M	224	External	75.73%	0.576
Yalniz et al. (2019)			Meta	74.06%	0.579
Inception-ResNet-V2	56M	299	External	78.23%	0.586
Szegedy et al. (2017)			Meta	78.25%	0.587
EfficientNet b4 NS	19M	380	External	83.66%	0.603
Xie et al. (2020)			Meta	84.85%	**0.630**
EfficientNet b5 NS	30M	456	External	78.25%	0,604
Xie et al. (2020)			Meta	85.94%	0.618
EfficientNet b6 NS	43M	528	External	85.99%	0.612
Xie et al. (2020)			Meta	86.07%	**0.630**
ConvNeXt-B	89M	384	External	85.91%	0.592
Liu et al. (2022b)			Meta	86.95%	0.594
CNNs Ensemble	**0.660**				

Table 6.8: 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 2019 Score	CNNs ensemble 2019 Score
Rank of probabilities	0.611	0.647
Majority voting	0.542	0.603
Averaging	**0.612**	**0.660**

Table 6.9: Ensemble method used for both the ViTs and CNNs

6.9 Model Selection

Once the previous results have achieved second place in the ISIC 2019 live leaderboard with the CNN ensemble, the best models to enhance the ensemble for ViT must be identified. The approach for determining the optimal ensemble is pro-

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 6.10: Outlier class metrics comparison with the OOD results for the top 1 in the 2019 ISIC live challenge

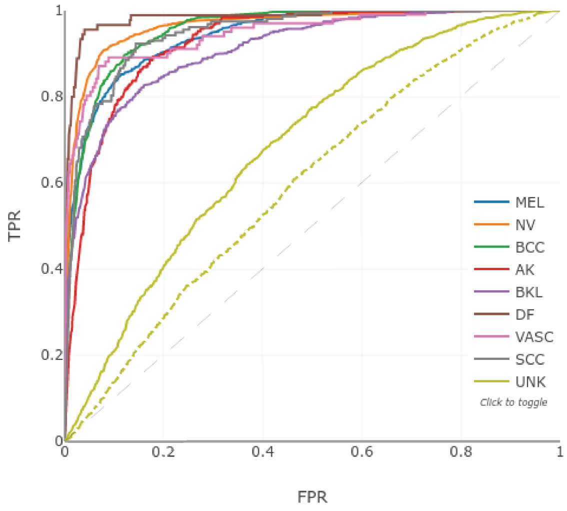


Figure 6.5: ROC curve with improvement AUC for the unknown class

vided here, which entails assessing a gap between models using the correlation of training with test predictions for each model. Therefore, MCM as used by Nikita Kozodoi (2020), was extended in this study for the nine class predictions (see Section 4.1.1).

Figure 6.6 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 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.

6.10 ViTs and CNNs Final Ensemble

Table 6.11 represents the ensemble that had reached first place in 2019 ISIC live challenge and third place in 2020 ISIC live challenge, see Figures 6.7 and 6.8. It was composed of a diversification of models, both ViTs and CNNs in Table 6.8, and

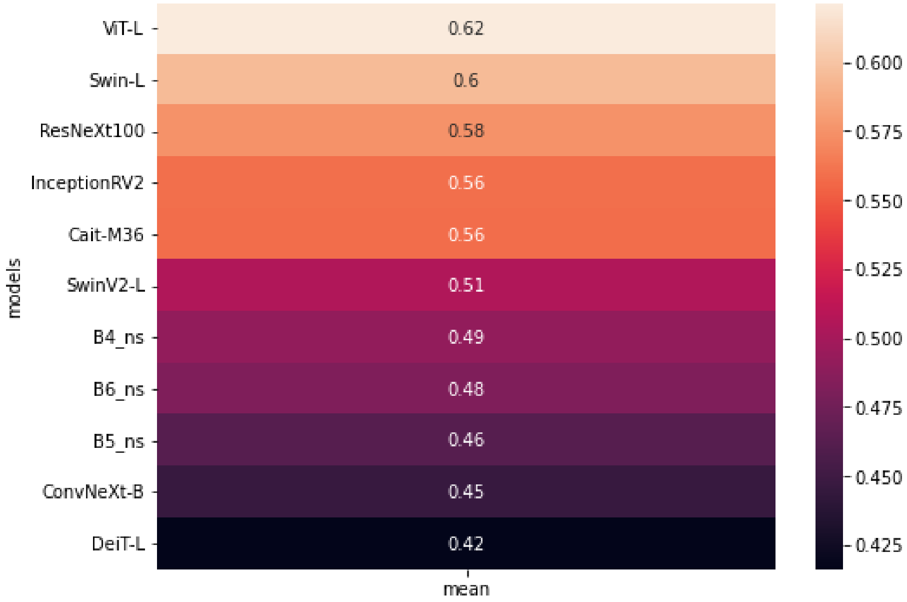
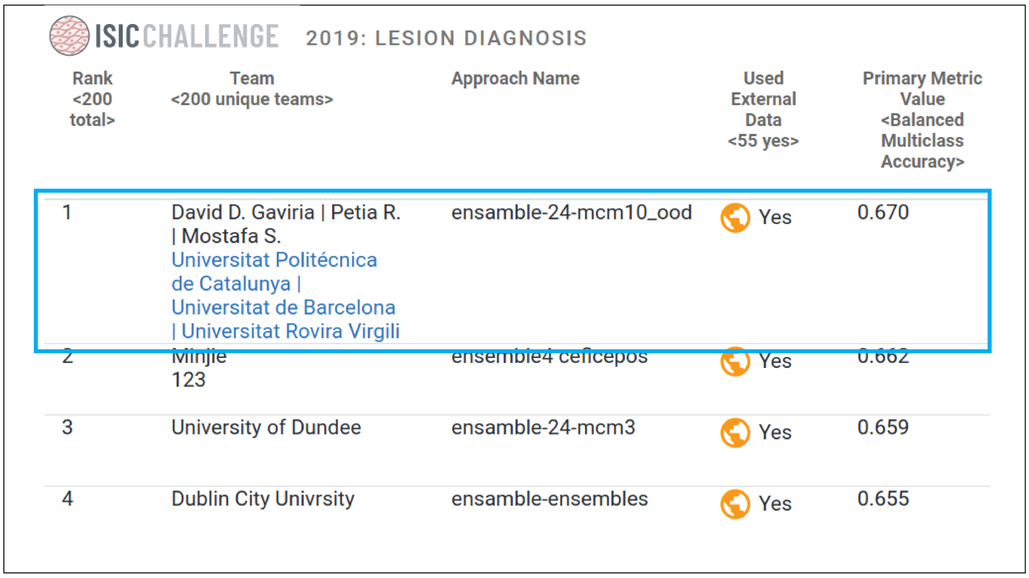


Figure 6.6: Mean correlation matrix of predictions for model selection. The higher gap means a poor model, likely overfitting on local data.

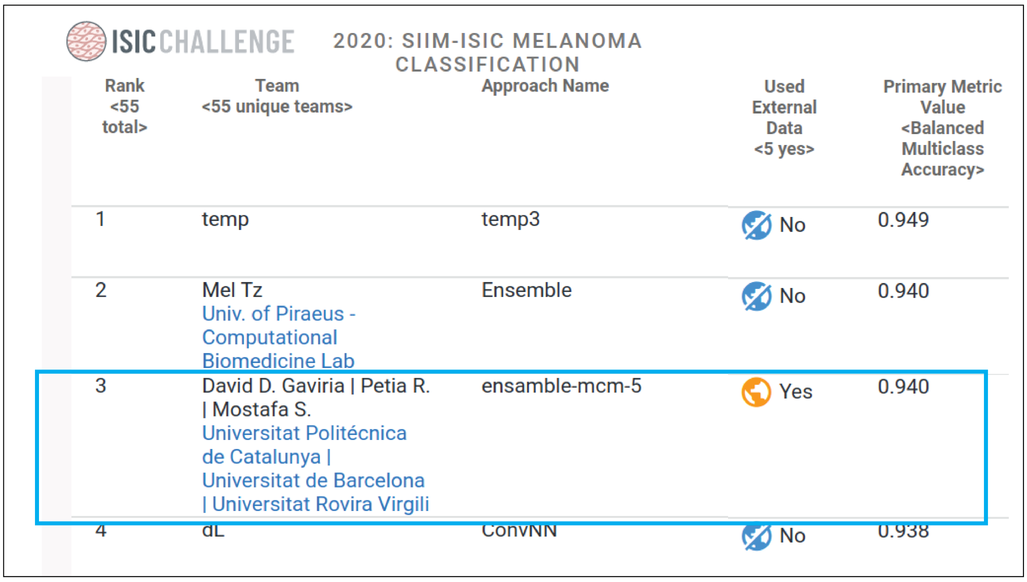
discriminated after a model selection with the MCM from the previous section 6.9.



ISIC CHALLENGE 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 cefcepos	Yes	0.662
3	University of Dundee	ensamble-24-mcm3	Yes	0.659
4	Dublin City Univrsity	ensamble-ensembles	Yes	0.655

Figure 6.7: First place in 2019 ISIC live leaderboard



ISIC CHALLENGE 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

Figure 6.8: First place in 2020 ISIC live leaderboard

6.10.1 ISIC Submissions and Evaluation

We submitted our model to the ISIC Challenge submission system, which allows for automatic format validation and scoring explained in 6.1. Figure 6.9 and Table 6.11 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 6.9 ROC curve and AUC reveals that the ROC curve performs much worse with the UNK class than with the other classes. Likewise from Table 6.11, 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 6.11 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 6.9 and the metrics results in Table 6.12 are the results of the submission to the 2020 ISIC live challenge. The 0.940 AUC allowed

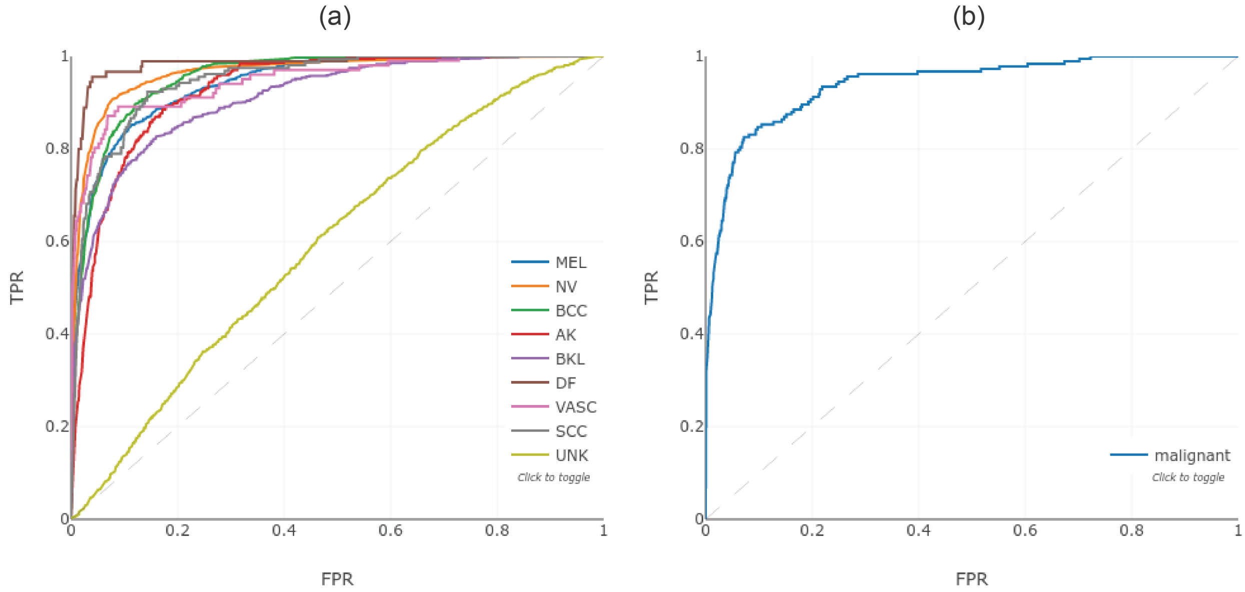


Figure 6.9: ROC curve for (a) the 0.670 balanced multi-class accuracy ensemble for the 2019 ISIC Challenge and (b) the melanoma with 0.940 AUC for the 2020 ISIC Challenge.

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 6.11: 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 6.12: Ensemble melanoma metrics for top-3 in the 2020 ISIC live challenge.

the project to finish third in the 2020 ISIC live challenge, confirming the proposal's generalization capabilities in a different test dataset.

Chapter 7

Discussions

In general, NNs have several millions parameters by which it can learn and seeks to optimize on those in a given domain. Since the search space of a non-trivial classification problem may be broad, various local minima for multiple initialization can be found. Similarly, even if the error and accuracy are comparable, models can rely on unique characteristics and hyperparameters to locate alternative local minima. This behavior explains the results in Section 6.8, demonstrating the effectiveness of ensembles at reducing the high variance observed on a single model.

It might be tempting to assume that just employing more complex or accurate backbones would improve performance across multiple competitions. However, our results suggest this is not always the case, especially in medical images, which are notorious for having unbalanced datasets with minority samples being the class of interest Gao et al. (2021), Qiao et al. (2021). Particularly for skin cancer classification, it was noticed that no single model can achieve the best results, because there are few positive cases in some of the classes, and the stochastic nature of NN plays a role. Overall, the stochastic nature of NNs explains why an ensemble is mandatory for such classification tasks, as well as an effective model selection strategy that ensures the diversification of a wide range of model architectures, in addition to acquiring contextual knowledge of the data and selecting a properly designed validation approach.

7.1 Validation Strategy

In order to establish a baseline performance on our proposed dataset and evaluate the performance of different features, we design experiments for two aspects:

(a) Comparing the influence of different backbones.

(b) Evaluating existing methods whose aim is medical skin image classification.

In all of the experiments, a stratifying split was made through the holdout method. 90% of the total stratified images with respect to the target as training, and the rest went to the validation set. First, an EDA was implemented to understand the data distribution and limitations. Following the data preprocessing, identification, removal and finally, the ensemble and its model selection approach were presented in full.

The hold-out method, which was used in the majority of the tests performed in this study, was covered as a straightforward train and validation split, with the cautious incorporation of a stratified split to keep the same target proportion of the original data for stability. Although, CV is commonly the preferred validation strategy for obtaining a stable model capable of learning from all the data, it is a computationally expensive strategy and especially with larger networks, it is unfeasible to conduct a large number of experiments, which were needed for the purpose of this work.

Our solution focuses from the start in understanding the data, from which dataleaks were found in the validation scheme, and thereby removing them in order to have more reliable validation metrics. Moreover, a variety of CNNs and ViTs were trained to address the skin lesion classification task. The models were all trained with the same settings, see 6.3.1, and the cropping approach was crucial to our score. Although DeiT is slightly below EfficientNet both in the literature and in our results, the ensemble shows that ViTs and ConvNeXts when used together can reach better results.

Furthermore, the pipeline built using OneCycle LR assisted the models in avoiding overfitting on the validation set with a short number of epochs, using transfer learning via fine-tuning, and networks performed best when no layers were frozen for any particular epoch of training. Furthermore, after training the models, TTA in conjunction with a the same random crop policy than training, but omitting CutOut improved performance when making predictions.

Overall, in order to increase the diversity of the ensembles, patient-level information and external data were integrated, and by monitoring the validation BACC metric, a variety of stable models were able to achieve high generalization performance. Finally, an extension of the Gram-OOD was used to compute the OOD and enhance the unknown class in order to solve the outlier problem in the test set. Finally, this validation approach aimed for reasonably comparing results that provide consistent results while avoiding overfitting on local data.

7.2 On Melanoma Diagnosis

Professionals typically evaluate the patient's unique "biologic skin ecosystem" when evaluating skin lesions for biopsy in relation to other lesions on the patient's body Rotemberg et al. (2021). For instance, an unusual lesion on a patient with other lesions that appear to be more benign is not considered to be as threatening as a lesion with malignancy-predictive traits amid numerous comparable lesions. This suggests using a patient-centric approach to design training and validation, which means utilizing a stratified splitting by patient and using the patient-id as a metadata input feature.

Other types of skin cancer rarely spread from outside of the surface of the skin, and the ability of the melanoma to methastasize makes it the most deathly. The good news is that first stage melanoma has a cure rate of about 95% so early detection is crucial. According to Rekha et al. (2021), 50% of patients have more than 10 contextual lesions, which justifies efforts to create a patient-centric dataset of images and metadata. Accordingly, this is done for the binary task of identifying particularly malignant melanoma using clinical context. However, it is important to note that this information was not used at all because, on the one hand, the overall goal was to assess the generalization capability of a nine target skin disease classifier in melanoma rather than to build a specialized binary classification pipeline focused solely on improving a binary classification with the AUC itself, which will give more weight to the malignant, but in detriment of the generalization gained form other skin lesions. On the other hand, the contextual-lesion information was not available in other datasets, except for the ISIC 2020 Challenge, limiting severely its exploitation potential.

To summarize, using diagnosis as target was proven effective by the ISIC 2020 Challenge winners in section 2.2.3.2, and a model concatenation of images and metadata was proven effective in combination with other models; contrasting previous work, the present study focused on a straightforward solution with an one-hot encoding of all metadata input and a simple data splitting for the image datasets. Without a doubt, incorporating patient-level contextual information has shown to aid in the creation of image analysis tools for clinical dermatology assistance, and therefore metadata were added into the pipeline, providing top results most networks, see 6. Finally, keep in mind that no computer-aided diagnosis is intended to replace medical expertise and experience, and that a human in the loop must always be included in the process and render the ultimate decision. However, the goal is to increase such software's generalization capability so that it can be more reliable at delivering diagnosis, avoiding overconfident predictions in the case of Out-of-distribution samples.

7.3 Societal Impact

The demand for GPUs, as in every DL classification project, translates into a high computational cost, which in the case of such competitions results in a broad consumption of energy to run the experiments. Over 300 models have been trained in this study, with an approximation of two models trained daily, running notebooks 24 hours a day. This accounts for a significant energy demand, particularly for models with large number of parameters. Nonetheless, on the one hand, the work provided is intended to mitigate this impact by providing the results in an accessible format that can be easily corroborated, and on the other, to contribute to skin cancer early detection and diagnosis through the usage of a non-existent ensemble of ViTs and CNNs submitted to the International Skin Cancer Detection Competitions.

7.4 ViTs vs CNNs Classification

Since the appearance of ViTs Dosovitskiy et al. (2021), CNNs and ViTs have enter into a race to see which one can outperform the other on the ImageNet Zhai et al. (2022), Liu et al. (2022a), and many sophisticated architectures, hyper-parameter tuning and data augmentation regimes have been proposed to improve the generalization. However, it has been proven that, for the time being, the unique characteristics of CNN with their strong sense of locality cannot be underestimated by the attention-based modules of the ViT architectures in real-world classification scenarios; namely, pixels' proximity, resemblance and color relationships, as well as the lack of convolution-like inductive biases challenging the ViTs Chen et al. (2022). Additionally, there have been studies as in Li (2021) addressing the robustness of OOD detection in ISIC images particularly with ViTs, and the limitations in that respect are an open book.

Chapter 8

Conclusions and Future Work

The work made on skin disease image classification is addressed in threefold:

(1) dermoscopic and clinical skin image classification with usage of patient-related metadata,

(2) an ensemble of a diversity of models, both ViTs and CNNs and

(3) the design of a carefully designed pipeline to assure low computational time and high generalization.

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 diversity, 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.

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 Yang et al. (2018). 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.

Additionally, while SAM with ViTs has been shown to beat CNNs without substantial data augmentations Chen et al. (2022), it requires n-fold computation time, which could not be afforded in this work and too, remains open for the future. Also, while thresholding gave a robust solution to the highly unbalanced dataset, contrastive Loss has produced promising results in a number of tasks, implying that future research should focus on evaluating cutting-edge Self-supervised Learning such as Simple Framework for Contrastive Learning (Sim-CLR) Chen et al. (2020), Simple Siamese Representation Learning (SimSiam) Chen and He (2021), and Nearest-Neighbor Contrastive Learning of visual Representations (NNCLR) Dwibedi et al. (2021), to improve melanoma prediction for the 2020 ISIC challenge.

Finally, we hope to be able to translate the findings here into other types of competitions which can be related to skin medical images such as Covid-19 diagnosis through chest radiographs SIIM (2021) or more general types of classification tasks such as food recognition with Food2K dataset Min et al. (2021), in order to further assess ViTs and CNNs generalization capability.

Appendix A

Augmentation

A.1 Samples of all the augmentation techniques tested

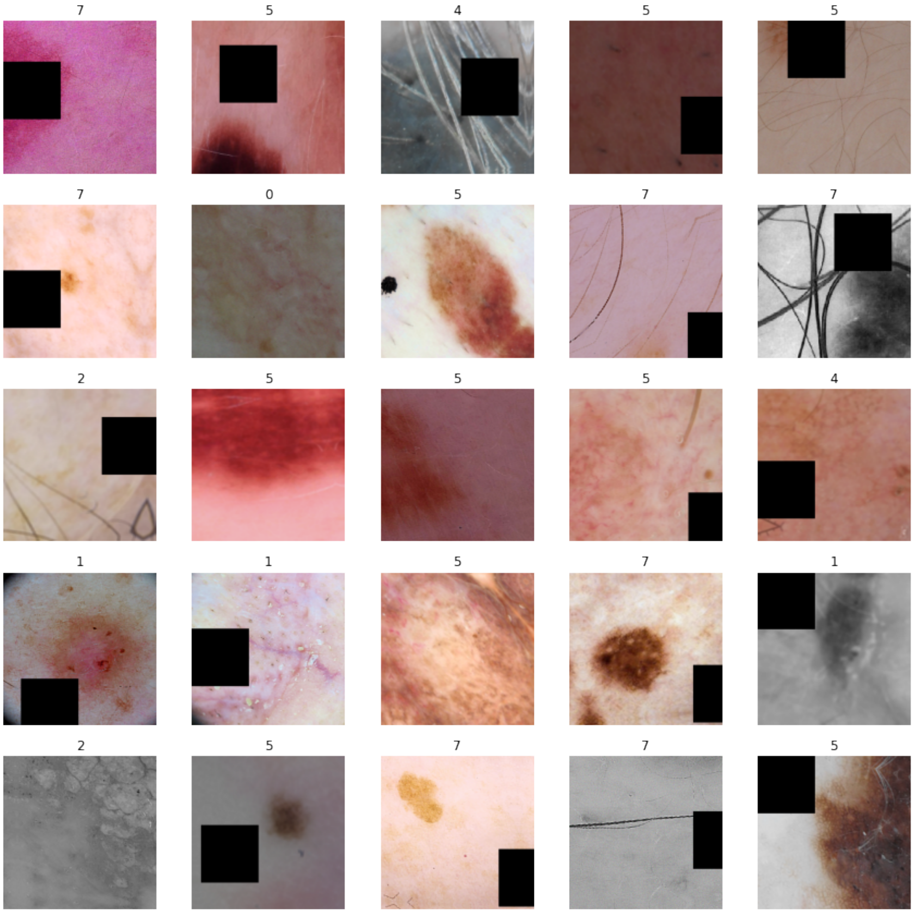


Figure A.1: Sample of final augmentation

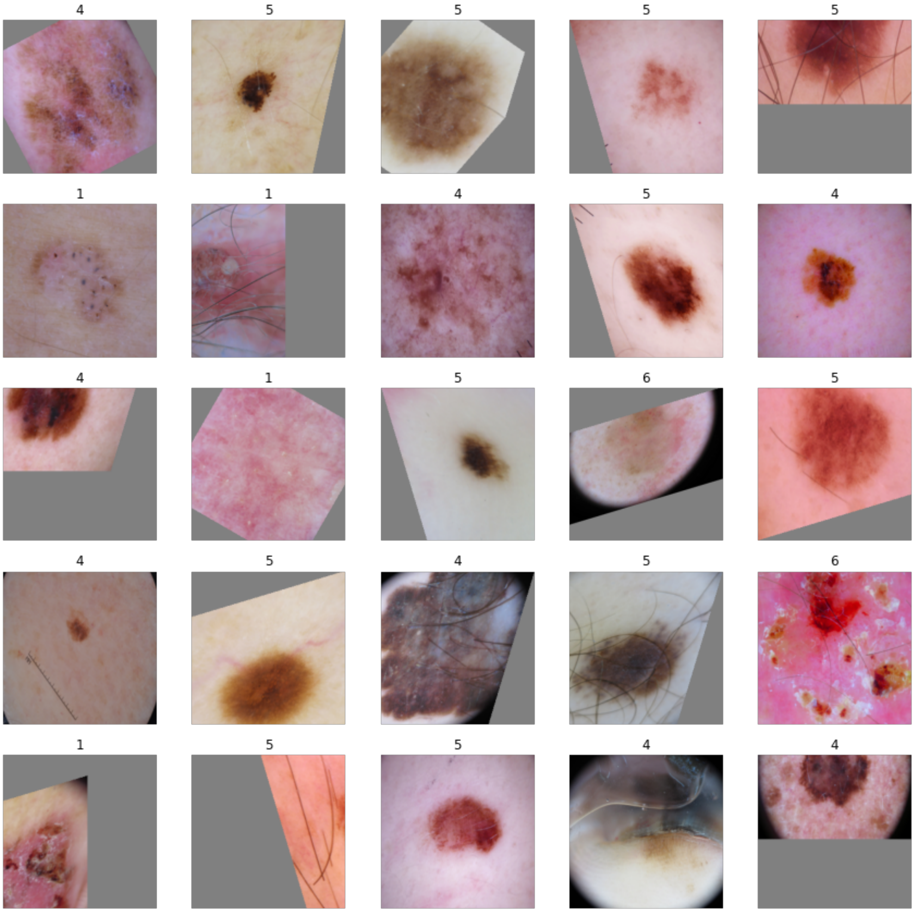


Figure A.2: Sample of RandAug tested on policy m9-n3-mstd0.5

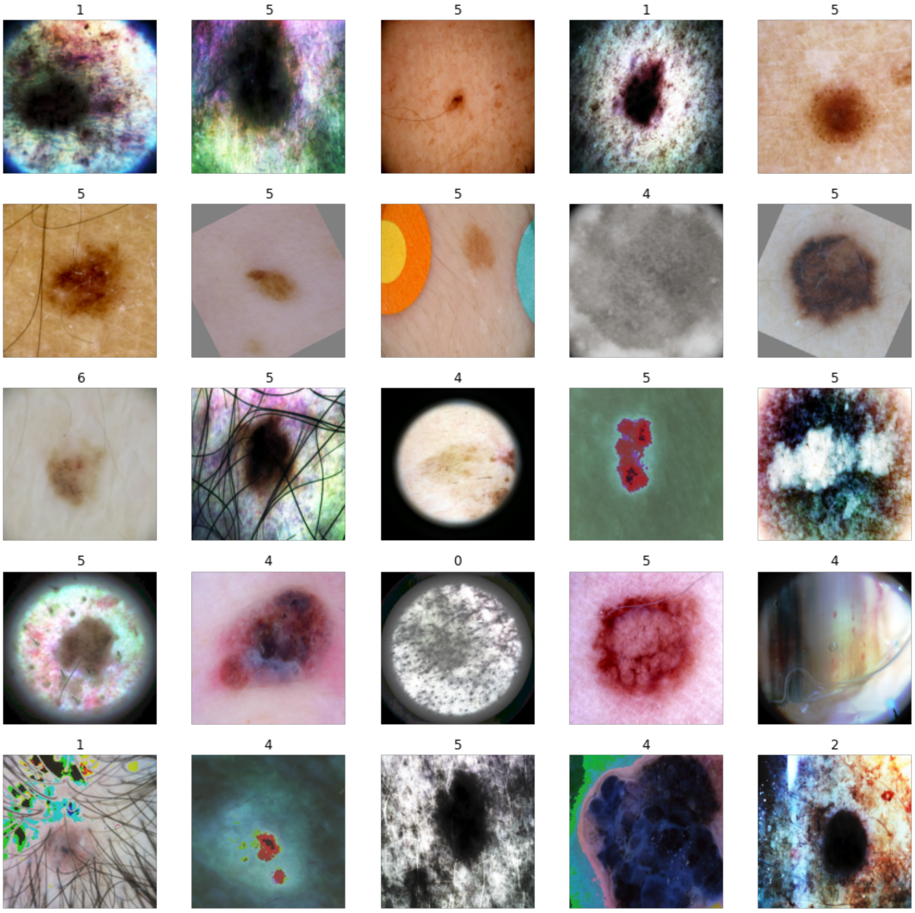


Figure A.3: Sample of AutoAugment tested on original-mstd0.5

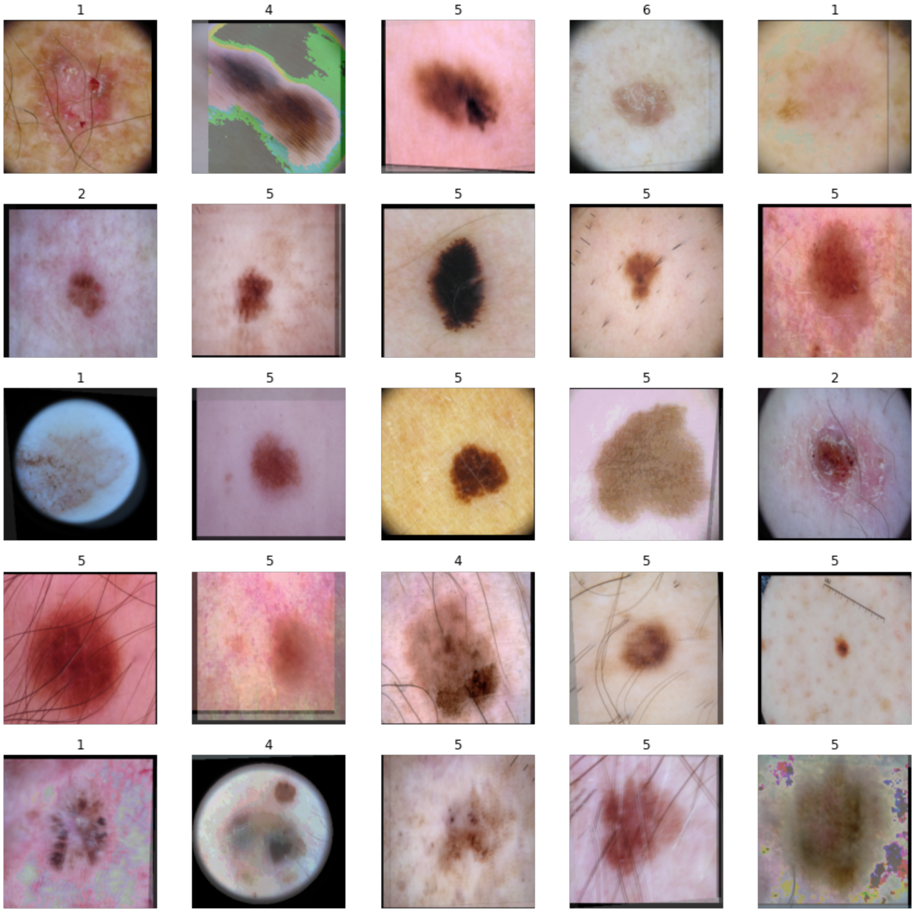


Figure A.4: Sample of AugMix tested on policy m5-w5-d2

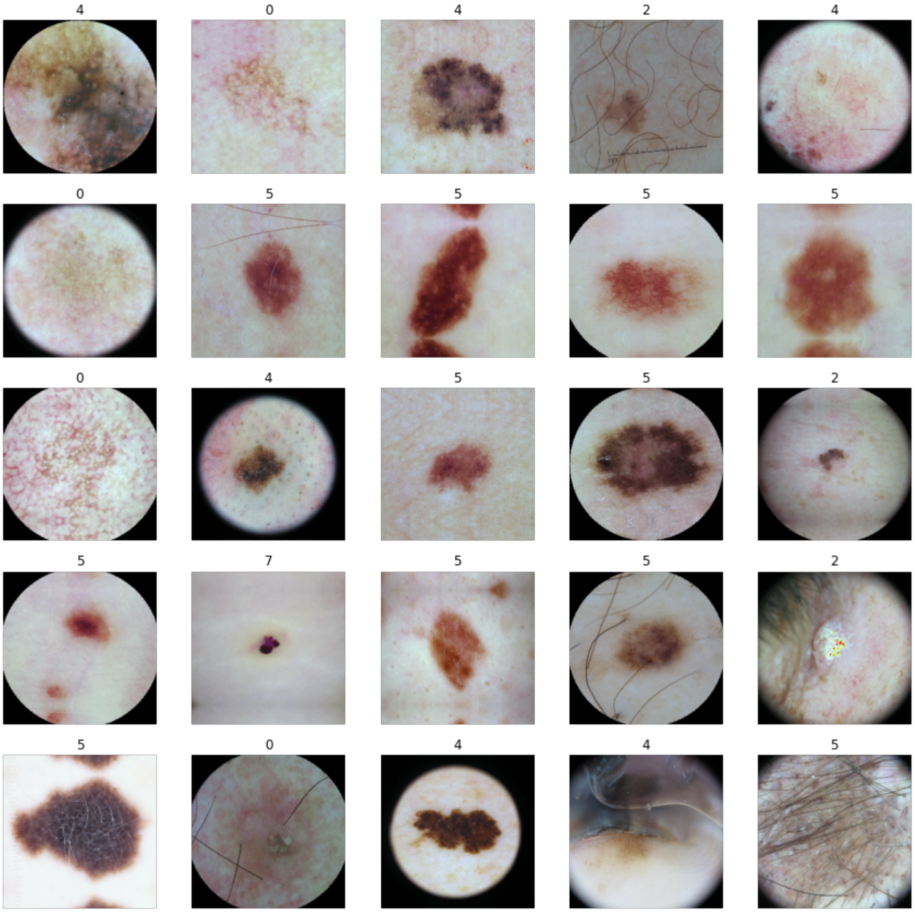


Figure A.5: Samples of the microscope-like cropping

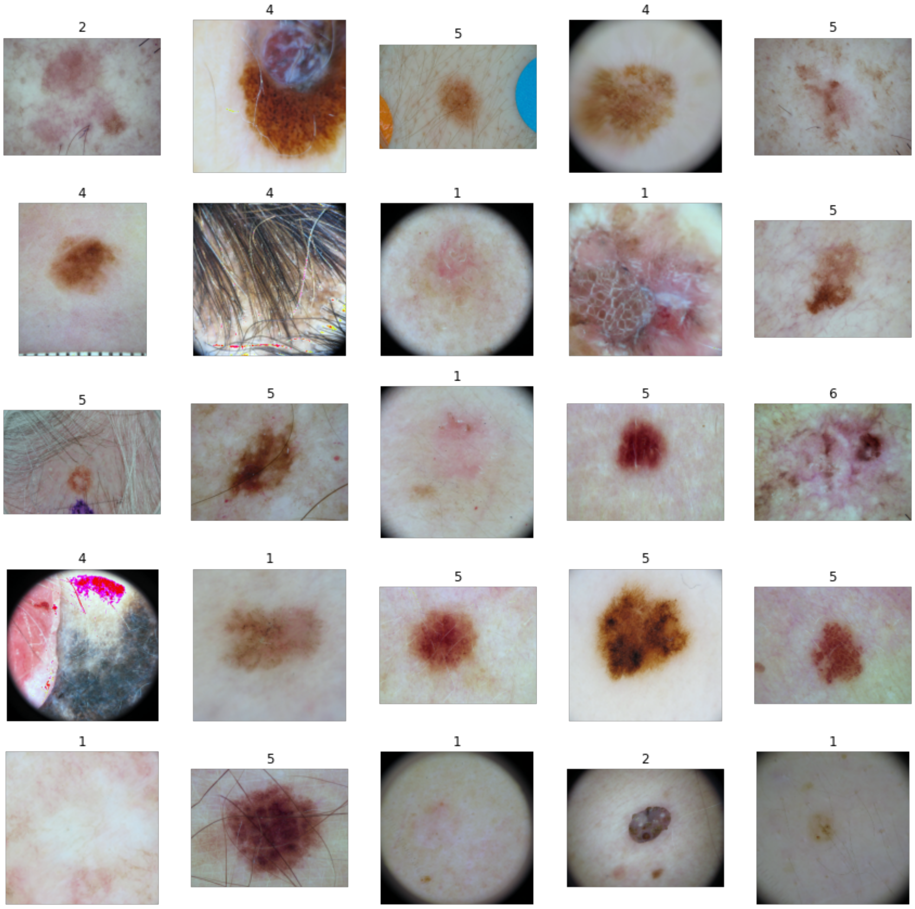


Figure A.6: Samples of the color constancy shades of grey technique

Bibliography

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

J. Ahmad, K. Muhammad, and S. W. Baik. Data augmentation-assisted deep learning of hand-drawn partially colored sketches for visual search. *PLoS ONE*, 12, 2017.

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

H. Alquran, I. A. Qasmieh, A. M. Alqudah, S. Alhammouri, E. Alawneh, A. Abughazaleh, and F. Hasayen. The melanoma skin cancer detection and classification using support vector machine. *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5, 2017.

A. American Cancer Society. Key statistics for melanoma skin cancer. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html. Accessed: 2022-08-30.

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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

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

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

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

M. S. Ayhan and P. Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. 2018.

Y. Bai, J. Mei, A. L. Yuille, and C. Xie. Are transformers more robust than cnns? In *NeurIPS*, 2021.

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

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

C. Berger, M. Paschali, B. Glocker, and K. Kamnitsas. Confidence-based out-of-distribution detection: A comparative study and analysis. In *UNSURE/PIPPI@MICCAI*, 2021.

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

P. J. Boland. Majority systems and the condorcet jury theorem. *The Statistician*, 38: 181–189, 1989.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 2004.

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

CDC.(2022). What is skin cancer? `https://www.cdc.gov/cancer/skin/basic_info/what-is-skin-cancer.htm`. Accessed: 2022-07-30.

M. E. Celebi. Dermoscopy image analysis in the age of deep learning. 2021.

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

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

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

X. Chen and K. He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021.

X. Chen, C.-J. Hsieh, and B. Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *ArXiv*, abs/2106.01548, 2022.

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

P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. J. C. Campilho. Towards adversarial retinal image synthesis. *ArXiv*, abs/1701.08974, 2017.

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

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

C. Deotte. Rapids cuml knn - find duplicates. `https://www.kaggle.com/code/cdeotte/rapids-cuml-knn-find-duplicates`, 2020. Accessed: 2022-06-30.

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

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

P. Dufour. How to correctly use test-time data augmentation to improve predictions. `https://stepup.ai/test_time_data_augmentation/`, 2020. Accessed: 2022-07-30.

D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9568–9577, 2021.

A. Esteva, B. Kuprel, R. A. Novoa, J. M. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

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

G. D. Finlayson and E. Trezzi. Shades of gray and colour constancy. In *Color Imaging Conference*, 2004.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *ArXiv*, abs/2010.01412, 2021.

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

S. C. Foundation. Skin cancer facts statistics. `https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/`, 2022. Accessed: 2022-08-30.

Y. Freund and R. E. Schapire. A short introduction to boosting. 1999.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 2004.

A. Galdran, A. Alvarez-Gila, M. I. Meyer, C. L. Saratxaga, T. Araújo, E. Garrote, G. Aresta, P. Costa, A. M. Mendonça, and A. J. C. Campilho. Data-driven color augmentation techniques for deep skin image analysis. *ArXiv*, abs/1703.03702, 2017.

L. Gao, C. Liu, D. Arefan, A. Panigrahy, M. L. Zuley, and S. Wu. Medical knowledge-guided deep learning for imbalanced medical image classification. *ArXiv*, abs/2111.10620, 2021.

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

I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov. Mednode: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Syst. Appl.*, 42:6578–6585, 2015.

J. Glaister. Automatic segmentation of skin lesions from dermatological photographs. 2013.

D. A. Gutman, N. C. F. Codella, M. E. Celebi, B. Helba, M. A. Marchetti, N. K. Mishra, and A. C. Halpern. 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)*, pages 168–172, 2018.

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

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

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

G. E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504 – 507, 2006.

S. I. C. ISIC-Colaboration.(2020). Overview of the isic collaboration. `https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/aboutIsicOverview`. Accessed: 2022-05-30.

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

J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23:538–546, 2019.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324, 1998.

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

X. Li. Out-of-distribution detection using vision transformers. 2021.

S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv: Learning*, 2018.

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

G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

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

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

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

A. Masood and A. Al-Jumaily. Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms. *International Journal of Biomedical Imaging*, 2013, 2013.

T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marçal, and J. Rozeira. Ph2 - a dermoscopic image database for research and benchmarking. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5437–5440, 2013.

W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang. Large scale visual food recognition. *ArXiv*, abs/2103.16107, 2021.

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

H. R. Narayanan. Automatic classification of skin cancer usingknn, svm and cnn. *Eurasian Journal of Analytical Chemistry*, 12:133–138, 2020.

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

U. of Iowa Health Care (2020). Malignant melanoma. `https://medicine.uiowa.edu/dermatology/malignant-melanoma`. Accessed: 2022-06-30.

K. O'Shea and R. Nash. An introduction to convolutional neural networks. *ArXiv*, abs/1511.08458, 2015.

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

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

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

F. Perez, C. N. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In *OR 2.0/CARE/CLIP/ISIC@MICCAI*, 2018.

T. Pircher, B. Pircher, E. Schlücker, and A. Feigenspan. The structure dilemma in biological and artificial neural networks. *Scientific Reports*, 11, 2021.

Z. Qiao, A. Bae, L. Glass, C. Xiao, and J. Sun. Flannel (focal loss based neural network ensemble) for covid-19 detection. *Journal of the American Medical Informatics Association : JAMIA*, 28:444 – 452, 2021.

O. Razeghi, G. Qiu, H. C. Williams, and K. S. Thomas. Skin lesion image recognition with computer vision and human in the loop. 2012.

C. Rekha, R. Jegatha, P. Sharmila, and A. Cibi. Skin lesion detection using convolutional neural network. 2021.

M. Rerábek and T. Ebrahimi. New light field image dataset. In *QoMEX 2016*, 2016.

M. D. Richard and R. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3:461–483, 1991.

L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33:1–39, 2009.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.

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

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

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

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

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks : the official journal of the International Neural Network Society*, 61:85–117, 2015.

D. Shanmugam, D. W. Blalock, G. Balakrishnan, and J. V. Guttag. When and why test-time augmentation works. *ArXiv*, abs/2011.11156, 2020.

S. Sharmeela and P. Asha. Classification of skin diseases by using back propagation neural network and abcd rule. 2017.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019.

R. SIIM, FISABIO. Siim-fisabio-rsna covid-19 detection. `https://www.kaggle.com/competitions/siim-covid19-detection`, 2021.

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

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

L. N. Smith and N. Topin. Super-convergence: very fast training of neural networks using large learning rates. In *Defense + Commercial Sensing*, 2019.

D. Steinkrau, P. Y. Simard, and I. Buck. Using gpus for machine learning algorithms. *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 1115–1120 Vol. 2, 2005.

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

X. Sun, J. Yang, M. Sun, and K. Wang. A benchmark for automatic visual classification of clinical skin disease images. In *ECCV*, 2016.

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

M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

A. Tejera-Vaquerizo, E. Nagore, J. J. Meléndez, N. López-Navarro, A. Martorell-Calatayud, E. Herrera-Acosta, V. Traves, C. S. Guillen, and E. Herrera-Ceballos. Chronology of metastasis in cutaneous melanoma: growth rate model. *The Journal of investigative dermatology*, 132 4:1215–21, 2012.

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

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

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

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

G. Vrbancic and V. Podgorelec. Transfer learning with adaptive fine-tuning. *IEEE Access*, 8:196197–196211, 2020.

F. M. Walter, A. T. Prevost, J. C. Vasconcelos, P. Hall, N. P. Burrows, H. C. Morris, A. L. Kinmonth, and J. D. Emery. Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 63 610:e345–53, 2013.

J. Weber(2020). 2020. siim-isic melanoma classification — kaggle post. `https://www.kaggle.com/competitions/siim-isic-melanoma-classification/discussion/161943`. Accessed: 2022-06-30.

K. R. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3:1–40, 2016.

W. H. O. WHO(2017). 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.

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

H. Wu, H. Yin, H. Chen, M. Sun, X. Liu, Y. Yu, Y. Tang, H. Long, B. Zhang, J. Zhang, Y. Zhou, Y. ping Li, G. Zhang, P. Zhang, Y. Zhan, J. Liao, S. Luo, R. Xiao, Y. Su, J. Zhao, F. Wang, J. Zhang, W. Zhang, J. Zhang, and Q. Lu. A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. *Annals of Translational Medicine*, 8, 2020.

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

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

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

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

J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *ArXiv*, abs/2110.11334, 2021.

C. yu Zhu, Y. Wang, H. Chen, K. Gao, C. Shu, J. Wang, L.-F. Yan, Y. Yang, F. ying Xie, and J. Liu. A deep learning based framework for diagnosing multiple skin diseases in a clinical environment. *Frontiers in Medicine*, 8, 2021.

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

C. Zauner. Implementation and benchmarking of perceptual image hash functions. 2010.

X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1204–1213, 2022.