TÉCNICO LISBOA

Skin Cancer Diagnosis Using Self-Supervised Learning

Maria Rita Ribeiro da Fonseca Verdelho

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Dra. Ana Catarina Fidalgo Barata
Prof. Jorge dos Santos Salvador Marques

Examination Committee

Chairperson: Prof. João Fernando Cardoso Silva Sequeira
Supervisor: Dra. Ana Catarina Fidalgo Barata
Member of the Committee: Prof. Pedro Manuel Quintas Aguiar

November 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to thank my parents, my two sisters and my brother for their support, encouragement and friendship throughout my academic career. I could not have come this far without all your help. I would like to especially thank my parents for giving me the opportunity and means to have the best academic education.

To my supervisors, Dra. Catarina Barata and Professor Jorge Marques, whose guidance, assistance and critical feedback were essential during this project execution. Thank you very much for all the help that allowed me to feel motivated and always supported.

A special thanks to Dra. Catarina Barata for all her constructive feedback, partnership and availability to help me write the article, that was based on this thesis, and was submitted for publication in the ISBI 2022 conference.

Lastly, I would like to thank my friends and remainder family for all their support during these 5 academic years.

Abstract

Convolutional Neural networks (CNNs) are the standard approach for image classification. However, they require a large amount of data and corresponding annotations. Collecting medical data is a difficult task, due to privacy restrictions. Moreover, it is even harder to obtain the clinical labels, since these must be provided by specialists. Self-supervised learning (SSL) has emerged as a possibility to overcome this issue, since it uses non-annotated data to pre-train the CNN. Recently SSL has been applied in the context of skin cancer. However, the results were not conclusive since a qualitative analysis was missing. Moreover, a proper analysis of the impact of different SSL approaches is still missing. In this master's thesis it will be investigated two SSL approaches: **Rotation** and **SimCLR**. The results highlight the benefits of applying self-supervised learning to the classification of dermoscopy images. Additionally, it was possible to demonstrate that these approaches learn different and complementary features, which is also a novelty of this thesis. As SSL is known to benefit from using more unlabeled data, it was also studied the impact of adding more data to the SSL pre-trained models (using 50% more data). It was possible to observe that depending on the level of difficulty of the task, the more it benefits from using more data. Therefore, the SimCLR task benefited more from the increase of data. The fusion of both techniques also showed to benefit with the use of more data, this was expected since the SimCLR also improved.

Keywords

Skin Cancer, Deep Learning, Self-Supervised Learning, Dermoscopy

Resumo

Redes neurais (CNNs) são a abordagem padrão para a classificação de imagens. No entanto, estes modelos exigem uma grande quantidade de dados anotados. O processo de obter dados clinicos é uma tarefa muito difícil, devido às restrições de privacidade que existem nos dias de hoje. Além disso, obter dados com as respetivas anotações médicas é ainda mais difícil de se conseguir, uma vez que estes diagnósticos têm que ser fornecidas por especialistas. A aprendizagem auto-supervisionada (SSL) surgiu como uma possibilidade de contornar este problema, na medida em que utiliza dados não anotados para pré-treinar as CNNs. Recentemente, o SSL foi aplicado no contexto do cancro da pele. No entanto, os resultados não foram conclusivos, sendo que lhes faltou uma análise qualitativa. Assim sendo, falta ainda executar um estudo onde se analisa o impacto das diferentes técnicas do SSL. Nesta dissertação de mestrado, foram investigadas duas abordagens de SSL: **Rotation** e **SimCLR**. Os resultados obtidos destacam os benefícios da aplicação da aprendizagem auto-supervisionada na classificação de imagens dermatoscópicas. Foi, também, demonstrado que essas abordagens aprendem recursos diferentes e complementares. O SSL é conhecido por melhorar o seu desempenho com o uso de mais dados. Consequentemente, optou-se por se executar uma experiência onde se adicionou mais 50% dos dados. Foi possível observar, que dependendo do nível de dificuldade da tarefa, mais esta beneficia do uso de mais dados. Assim sendo, o modelo pre-treinado com o SimCLR beneficiou mais com o aumento de dados. A fusão das duas técnicas também mostrou benefícios, o que era espectável, uma vez que o SimCLR também melhorou.

Palavras Chave

Lesões de Pele, Redes Neuronais, Aprendizagem Auto-Supervisionada, Imagens Dermoscópicas

Contents

List of Figures

List of Tables

Acronyms

Adam	Adaptive Moment Estimation
AKIEC	Actinic Keratosis
AUC	Area Under the Curve
BACC	Balanced Accuracy
BCC	Basal Cell Carcinoma
BKL	Benign Keratosis
CNN	Convolutional Neural Network
DF	Dermatofibroma
FCL	Fully-Connected Layer
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
ISBI	IEEE International Symposium on Biomedical Imaging
ISIC	International Skin Image Collaboration
LIME	Local Interpretable Model-agnostic Explanations
MEL	Melanoma
MLP	Multi-layer perceptron
NV	Melanocytic Nevus
ReLU	Rectified Linear Unit
SCC	Squamous Cell Carcinoma
SE	Sensibility

SSL	Self-Supervised learning
SP	Specificity
SVM	Support Vector Machine
TL	Transfer Learning
TN	True Negative
TP	True Positive
VASC	Vascular

1

Introduction

Contents

1.1 Motivation

Skin cancer is one of the most common types of cancer worldwide [18]. The main known cause is excessive exposure to the sun's UV rays, which over time, penetrates and damages the skin. Therefore, malignant lesions are likely to appear in areas that are more exposed to radiation such as limbs, back, face and neck [18].

Each year there are approximately 13.000 new cases of skin cancer in Portugal [18] and in the U.S. more than 9.500 people are diagnosed every day [19]. More people are diagnosed each year with skin cancer, in the U.S., than all other cancers combined. In the past decade, the number of melanoma cases diagnosed annually has increased by 47% and about 86% of these cases can be attributed to exposure to ultraviolet radiation [19]. In non-melanoma cancer about 90% of skin cancer is associated with UV radiation and about 5.400 people worldwide die every month due to this disease [19].

Skin cancer is also one of the most treatable forms of cancer when detected in an early stage. Also, late detection can have a significant impact in mortality rates [20]. Therefore, there is a need to develop a convenient and precise method to perform earlier diagnose and detect skin cancer lesions. However, this detection is not easy since the different lesions have many shapes, textures and colors that can be visually similar among melanoma and non-melanoma lesions [20]. Over the past decade, automatic methods based on deep learning have been developed to assist human experts and accelerate the process of cancer diagnoses.

There is a continuous need to improve the performance of the developed deep learning methods, to achieve a faithful classification of the different skin lesions. However, obtaining satisfactory results requires a huge amount of data, which is a very difficult task. Not only due to privacy and law restrictions but also because obtaining clinical labeled data requires the knowledge of a specialist. This problem is addressed in the section 1.3 and is one of the main focuses of this thesis.

1.2 Skin Lesions Analysis

Dermoscopy is a non-invasive diagnostic tool that dermatologists use to evaluate skin lesions regarding colors and micro-structures that are invisible to the naked eye. The main principle of this tool is to place fluid on the lesion, because oily skin allows light to pass through it and to reach the deeper dermis and consequently visualize subsurface skin structures [21]. After placing the fluid on the skin, a dermatoscope is used to magnify the lesion and the doctor is able to inspect it in more detail [21]. However, even though the lesion is magnified it is still a very hard task, even for experienced doctors, to diagnose the different lesions [21].

Dermatologists divide skin lesions into two main classifications, non-melanocytic and melanocytic groups, after this differentiation, they classify the lesions in benign or malignant classes. This hierarchy

is described in figure 1.1.

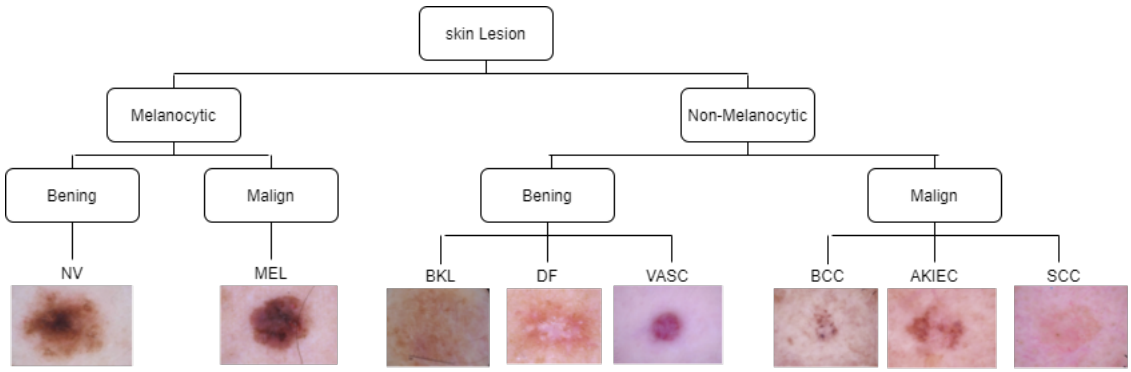


Figure 1.1: Skin lesions taxonomy of International Skin Image Collaboration (ISIC) 2019 dataset (Dermoscopy images extracted from [1] [2] [3]).

1.3 Problem Formulation

1.3.1 Problem Statement

Despite the great advances in developing methods to assist human experts and accelerate the process of cancer diagnoses, the existing methodologies are still far from allowing a robust performance of classifiers based on neural networks.

Deep neural networks are the most used methods for image classification. However, when applied to medical image analysis, the use of such techniques becomes challenging. These methods require huge amounts of data and corresponding labels, in order to achieve satisfactory and generalizable results [22]. Collecting clinical data is a difficult task, due to privacy and law restrictions. However, it is even harder to obtain clinical labeled data, since the labels must be provided by specialists. Therefore, the process of creating labels is very expensive and requires too much time, which doctors do not have, to get an acceptable number of annotated images [23] [24]. On the other hand, collecting unlabeled data is easier [25]. Although there has been an attempt of solving this issue by developing semi-automatic software tools that generate labels, these techniques resulted in having a non-significant impact on reducing the time spent for annotating these datasets [26].

In order to reduce the use of an excessive number of labeled images, current deep learning methods use Transfer Learning (TL). This method consists first in training a model for a task in a large data base (e.g., ImageNet[1]) and then the model is 'recycled' for a new target task, for example, a more specific image classification task, such as skin cancer diagnosis, or a different task (e.g., object detection) [28]. These pre-trained models on ImageNet usually have deeper architectures than what is needed in medical image analysis [29]. The distribution intensity from the natural images (images from ImageNet [27]),

[1] ImageNet [27] is labeled image database with more than 100,000 images.

is also very different in comparison to medical images [25]. Therefore, when trying to apply the previous knowledge obtained using natural images to the medical images, there are neurons, of the network, that remain loyal to the dataset, which can have difficulties in generalizing to the other data [29]. In fact, when using the unrelated dataset there is no need to have medical knowledge to label the images, but it is still required to use a significant amount of labels in this initialization step [23]. Therefore, the following question arises: Is it possible to learn weights for the medical image domain using fewer annotations and only images from the same domain?

1.3.2 What is Self-Supervised learning (SSL)?

A technique has been used to avoid the need to use huge amounts of labeled data in order to achieve satisfactory and generalizable results [22]. SSL was created to optimize the data usage, since this technique does not require the use of labels in the pre-training phase [24]. Its fundamental idea was inspired by how humans learn different tasks. First, it is essential to have a clear representation of the world, and only then a task can be learned. Babies, before learning how to walk, start to experience gravity, as well as start to understand the need to avoid objects and observe how other humans walk [30]. Therefore, during their life, humans learn tasks by observing their surroundings [30].

SSL, in contrast to TL, does not require the use of annotations by learning images representations from its pixels [31]. This is advantageous in the medical image domain, not only because there are a higher number of non-annotated datasets than annotated [23], but also because TL uses natural images (e.g., from ImageNet) which have significantly different properties from medical ones. Further details about the differences between TL and SSL are presented in section 2.3.1.

In fact, by applying SSL the neural network learns features that well represent the data and this knowledge can be very advantageous when applied to different tasks. So the question that now arises is: **How does SSL work?**

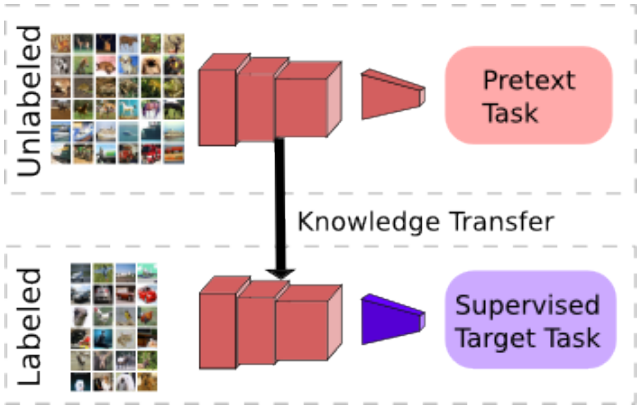


Figure 1.2: The main idea of SSL (extracted from [4])

Figure 1.2 summarizes the concept of SSL. The first step of SSL consists in assigning a simple task, known as pretext task (selected according to the intended goal) for the Convolutional Neural Network (CNN) to solve. There are multiple pretext tasks, which will be discussed in section 2.3.4. SSL is focused on the knowledge of the learned features rather than the final performance of the pretext task. In this first step, the network can learn a new task by learning from the unlabeled data. During this process, visual features from the unlabeled images are extracted and the CNN is trained to minimize the objective functions. The corresponding learned weights will, in the second step, be transferred to the target task. This consists of 'recycling' the first convolutional layers that contain the learned weights and applying that network to the target task by adjusting its final layers according to the intended goal. This final step consists of training a labeled dataset (with fewer annotated data) on the target task using the knowledge obtained by the pretext task. The final task is also known as target task and consists of image classification, segmentation, object detection, action recognition or others [25].

1.4 Thesis Objective and Contributions

SSL is gaining popularity as it is achieving promising results when comparing with TL [32]. Therefore, this thesis's main goal is to apply SSL techniques to the skin cancer diagnosis in order to better exploit the unlabeled images and improve the performance of classifiers based on neural networks. By doing so it is expected to prevent the use of huge amounts of annotated data. This thesis will combine TL with SSL in an attempt to filter the generalization problem that occurs when using TL. TL uses natural images that have a different domain to the skin lesion ones. Therefore the network resulted from applying TL, will have neurons that remain loyal to the natural images. By applying SSL it is expected to correct these neurons and obtain a network that generalizes better to the skin lesion images. In this thesis, it was performed a systematic assessment of six TL pipelines (two with supervised learning as the baseline and four self-supervised contenders) in 5 different partitions of the ISIC 2019 dataset. To better exploit the results it was executed a quantitative and qualitative analysis of each model. This is believed to be the first work that provides a qualitative analysis of the features learned by the SSL strategies.

From this thesis, an article has been submitted for publication in IEEE International Symposium on Biomedical Imaging (ISBI) [2] 2022 conference. The article is in Appendix B.

[2] ISBI is a scientific conference dedicated to mathematical, algorithmic, and computational aspects of biological and biomedical imaging, across all scales of observation. It fosters knowledge transfer among different imaging communities and contributes to an integrative approach to biomedical imaging.

1.5 Organization of the Document

This report is organized as follows: chapter 1 provides an introduction to the skin cancer detection, which involves a huge amount of data and an introduction to the SSL; chapter 2 addresses a brief description of CNNs, followed by a description of the existent state of the art, addressing the supervised and self-supervised methods used for detecting skin cancer and contains also a set of techniques used for self-supervised models; Chapter 3 explains this thesis approach with a more detailed explanation of the used methodologies; Chapter 4 starts with a dataset description, followed by a description of the evaluation metrics, the computational environment and the experimental results and discussion of the different initialization techniques developed during this thesis; Finally, chapter 5, contains the conclusions and possible future work.

2

Background

Contents

This chapter contains a brief explanation of the background of deep learning. First, starts with a general description of CNNs. Secondly, is followed by an explanation of the supervised learning technique, which is currently the most common technique used in skin cancer diagnoses, and thirdly the differences between SSL and TL are also stated. This chapter also contains the previous works in skin lesions diagnosis and ends with a description of SSL techniques and their characteristics.

2.1 CNNs

As mentioned in chapter 1 one of the most popular deep learning techniques are the CNNs. These networks have many different applications from image recognition to image classification or object detection among others [22]. As the name indicates, CNNs have architectures that were inspired by the human brain neurons.

CNNs receive images as input and assign different importance values, given by learnable weights and biases, to multiple objects in the image. These parameters allow the network to distinguish different images [33]. By applying different filters to the image, the network is capable of capturing the temporal and spatial dependencies within an image [33].

2.1.1 Basic Concepts

CNN receives images as inputs and submits them to a series of convolutional layers with filters, also known as kernels, followed by a non-linear activation function, in order to extract features from the images. The output of each layer is known as the feature map, which consists of an image different from the original. The feature maps will be submitted to a pooling layer that allows the CNN to reduce the dimension of each image. This process is repeated as many times as needed. The first convolutional layer extracts low-level features from the input image, e.g. edges and color, among others and the other convolutional layers allow the system to learn high-level features, which aim to a better understanding of the different images [33].

Finally, a Fully-Connected Layer (FCL) is applied to convert the feature maps into a single array. This helps to reduce the computational complexity required to process the data [33]. The set of FCLs is known as softmax classifier, which performs the intended multi-class classification by assigning a probability of each class label over all the classes [34]. Figure 2.1 presents a typical CNN architecture.

The structure of the CNN involves many hyper-parameters (variables that influence the structure of the network) that are directly related to the network efficiency [34].

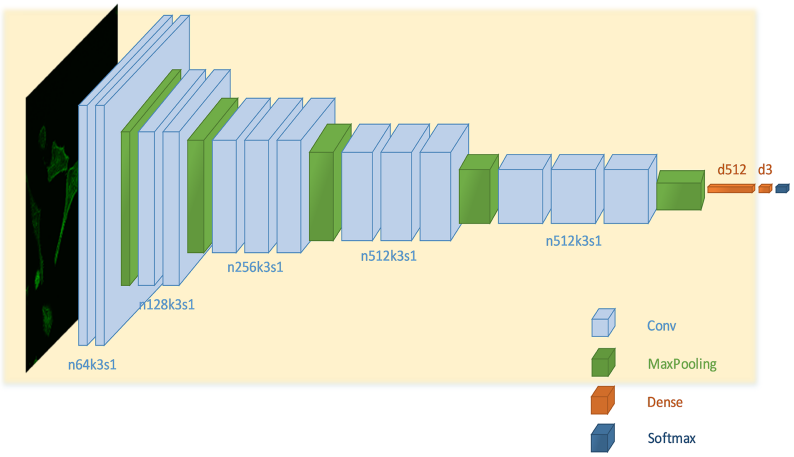


Figure 2.1: CNN architecture (extract from [5])

2.1.2 Training the model

The training phase of the CNN has the aim of optimizing the model's weights, to allow the network to better map the input to the correct predicted class [33]. A loss function is used to improve the quality of these output predictions by comparing the predicted output to the true label. Many different loss functions have different objectives.

The training phase can be seen as an optimization problem, where the minimum of the loss function is being searched. The network parameters are optimized through the gradient descent method, which indicates the right direction for the next iteration, in order to achieve the minimum of the loss function.

There are two phases when training the network. The forward phase, where the input goes through the network and the backward phase, where the gradients are back propagated and the weights are updated [33]. The latter phase is where the gradient of the loss function is calculated.

The weights initialization is a hyper-parameter of the network. The choice of this initialization is typically done, in supervised learning, by training the network from scratch or by using TL with pre-trained models [34]. Section 2.2 contains a more detailed description of supervised learning. By computing the forward phase an output is obtained and a loss is computed. The back propagation phase initiates and the gradient of the obtained loss function is computed. To reduce the loss function value the weights are updated. However, as a new model is being trained from scratch there is a need to use huge amounts of data.

2.1.3 ResNet Architecture

All the experiments carried out in this thesis use the ResNet-50 architecture. Thus, a brief overview of this network will be provided. The work presented in [6] introduces the concept of a residual neural network that aims to facilitate the training of deep neural networks. In the past, it was proven that with the

increase in the depth, the accuracy of the model tends to saturate and, then, degrades rapidly. In other words, by adding more layers into a previously trained network there is a decrease in the accuracy of the model. To avoid this problem, instead of staking layers directly, this paper proposes a novel solution that consists of replacing the traditional convolution blocks with residual connections. Figure 2.2 shows the pipeline of the residual blocks. These residual connections can be seen as 'shortcuts' that can be directly used once the input and the output have the same length.

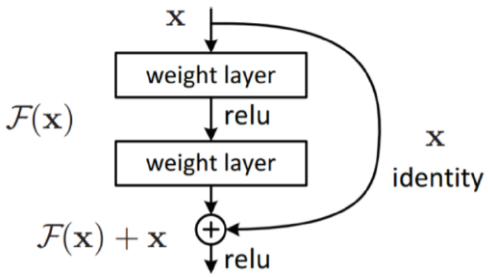


Figure 2.2: Pipeline of a Residual block (extracted from [6]).

As ResNet has proved to be a less complex network and nevertheless it still manages to obtain good results this is why it will be used during this thesis work.

2.2 Supervised Learning

Over the last three decades there has been an effort towards the development of machine learning methods to detect and classify the different skin cancer lesions. These methods are being created in order to help dermatologists correctly diagnose the different lesions [24]. In the past, instead of CNNs, was used Support Vector Machine (SVM), K-Nearest Neighbor, forest tree classifiers among others [35]. However, nowadays CNNs are the ones achieving the best results [22]. This can be seen by the type of works used for the International Skin Imaging Collaboration Challenge, also known as ISIC [36]. The question that now arises is: What are the most popular methods to diagnose skin cancer lesions?

The most common approach in deep learning is to apply supervised learning on neural networks [37] [32]. In supervised learning, the network is given a labeled dataset and is trained to predict the output of the corresponding input image. Therefore, when training the neural network, the system is capable of learning from experience [37]. During the training phase, the network learns to extract discriminative features from the images [38]. The accuracy of the system depends on both the amount and quality of the available data and the used architecture.

In the context of skin cancer, supervised learning is also the most common approach. In order to classify different skin lesions, the images from the dataset are considered as features and the medical annotations associated with each image are the labels. The network is trained using labeled images

and it is expected to acquire knowledge from the dataset in order to generalize the learned information to the new input images [35]. This confirms that the main supervised learning problem resides in the collection of a large dataset [39]. However, these large amounts of data are not easy to collect. Obtaining medically labeled images is even harder [20].

In supervised learning, the choice of the weights initialization is typically done by training the network from scratch or by using TL with pre-trained models. Training from scratch resides in assigning arbitrary weight values to the system, which means a new model is being constructed. On the other hand, TL uses weights that have already some image knowledge [28]. Therefore, this technique avoids the use of huge data, which resulted in an easier and faster method when compared to training the network from scratch [34].

2.2.1 TL

TL, as the name indicates, uses the foundation of exporting knowledge from one task to another. This technique uses a model already pre-trained in a labeled dataset and 'recycles' some of the initial convolutional layers that have acquired some knowledge and train the rest of the layers to adjust to the new target task. This way, the network begins with weights that have already some image knowledge that has little similarities to the medical images. There are two phases when using TL: pre-training and the fine-tuning phase.

The pre-training phase consists in training the network in a bigger and different dataset (e.g. ImageNet), which prevents overfitting. The model is gaining general knowledge since it is being forced to learn new images. Therefore, this results in a network that learned better representations from the images [32]. A common approach is to use pre-trained CNNs, which have been trained using datasets containing huge amounts of data. The main goal of this phase is to generalize and export knowledge to a new target task. Mainly, instead of starting the target supervised task with no knowledge, the key idea is to start already with some information that could be used to obtain better results in the performance of the system.

In the fine-tuning phase, the weights are then transferred to the target task. The most common approach is to modify the last FCLs of the network and train the new model to the target task, for example, image classification [35]. Most works using deep neural networks for skin lesion detection either use TL [27] or train them from scratch [28]. There are a set of different architectures used in TL. Being the most commonly used the AlexNet [40], Google Inception V3 [41], ResNet-50 [6], Xception [42], VGG-19 and VGG-16 [43].

TL can be very advantageous. However, once this technique requires the use of a non-related dataset the learned weights can have problems generalizing well enough to the target tasks and datasets. Since the classes from both tasks are very different [32]. Another visible limitation is the fact that there

is still the need to use labels in the pre-training phase.

2.2.2 Skin Cancer Diagnosis

Table 2.1 shows different works that applied supervised learning techniques to skin cancer image analysis using the ISIC 2019 and 2018 challenge datasets [1] [2] [3]. For the ISIC 2018 challenge three works ([44] [45] [46]) from the top-7 leadership have been selected and for the ISIC 2019 three works ([47] [48] [49]) from the top-5 have been selected. Table 2.1 contains the following metrics: Specificity (SP) and accuracy (ACC).

Table 2.1: Application of supervised learning to skin cancer diagnoses using ISIC challenge dataset [1] [2] [3].

Dataset	Authors	Extra Data	Arquictecture	Tranf. Learning	SP(%)	ACC(%)
ISIC 2019	Zhou et al., 2019 [47]	-	Ensemble	Yes	95,2	91,7
	Pollastri et al., 2019 [48]	-	Ensemble	Yes	96,3	92,4
	Chouhan, 2019 [49]	Private	DenseNet	Yes	96,7	91,0
ISIC 2018	Gessert et al., 2018 [44]	HAM10000 dataset	Ensemble	Yes	98,4	97,2
	Li et al., 2018 [45]	-	ResNet50	Yes	97,6	96,9
	Pan et al., 2018 [46]	-	ResNet	Yes	96,7	95,9

From analyzing table 2.1 it is possible to verify that TL is a common approach used for skin cancer diagnostic.

Thus, other questions arise: what if we could join the supervised learning technique with the self-supervised, which is currently gaining popularity, and apply it to skin cancer diagnoses? This could solve the TL limitations.

2.3 SSL

At this point, it is known that supervised learning, based on CNNs, should require the use of large amounts of labels. However, there are a huge number of images that are not labeled and that can be used by neural networks. As mentioned before, collecting manual labels is a very expensive and time-consuming task, since the annotation is required to be done by an expert [23] [25].

Therefore, the concept of SSL emerged as there was a need to get advantage of the available unlabeled data without creating labels, while still extracting visual features from the images.

2.3.1 TL vs SSL

SSL is similar to TL, but instead of pre-training a network using a labeled dataset, it uses an unlabeled dataset and extracts feature representations from the images by forcing the network to execute simple tasks. While executing these simple tasks the network learns parameters that are fine-tuned on the

target task. In other words, the weights obtained during the visual feature extraction phase are then used to initialize the convolutional layers of the CNN. Therefore, SSL recycles the first convolutional layers of the pre-trained network (trained on the unlabeled dataset) and adjusts the rest of the layers to the new target task.

2.3.2 Combining TL with SSL

The most common approach, while using SSL is to combine it with TL. Therefore, first the ImageNet weights are used and, then, they are redefined using SSL, in an attempt to filter the generalization problem that occurs when using only TL.

Recently, some works have adopted self-supervised approaches in the context of medical image analysis. In the appendix A, table A.1 is presented. It shows different works that applied SSL techniques to different medical applications. However, it is important to stress that most self-supervised techniques are very recent and, consequently, there are still few works that use them. The main question that now arises is: Does it make sense to apply SSL to the skin cancer diagnosis? This question is addressed below.

2.3.3 Skin Cancer Diagnostic

SSL is a relatively new concept and consequently there are very few papers applied to the skin image analysis. Table 2.2 shows different works that applied SSL techniques to different skin cancer problems.

Table 2.2: Application of SSL to skin lesion diagnosis.

Authors	Goal	Features	Score	SSL	TL	From Scratch	TL + SSL
Li et al., 2020 [24]	Segmentation	ColorMe	Dice (%)	86,3	86,7	84,6	87,7
Tajbakhsh et al., 2019 [29]	Segmentation	Colorization	Acc (%)	35	52	33	-
Kwasigroch et al., 2020 [23]	Classification	Jigsaw	AUC (%)	-	82,5	-	83,4
		Rotation		-			84,2
Chaves et al., 2021 [50]	Classification	BYOL	AUC (%)	-	94,8 $\pm$ 0,6	-	94,6 $\pm$ 0,5
		InfoMin					94,4 $\pm$ 0,5
		MoCo					93,9 $\pm$ 0,7
		SimCLR					**95,6 $\pm$ 0,3**
		SwAV					95,3 $\pm$ 0,6

Analyzing table 2.2 it is possible to conclude that the application of SSL to skin cancer diagnoses can lead to better performance when combined with TL [24] [23] [29] [50] (last column).

Both Li *et al.* [24] and Tajbakhsh *et al.* [29] applied SSL techniques related to color to the segmentation of skin cancer. Kwasigroch *et al.* [23] applied two SSL techniques related to geometric distortion to the skin cancer classification task. The closest work to this thesis is a preprint by Chaves *et al.* [50], in which they assess five self-supervision learning candidates using contrastive techniques against a competitive supervised baseline and conclude that SSL is competitive both in reducing variability and

improving accuracies. However, all works lack of a qualitative assessment of the impact of the different pre-training strategies.

The different SSL techniques will be addressed below.

2.3.4 SSL Techniques

SSL consists of two phases: the pre-training phase and the target task (which in this case is the image classification). In the pre-training phase, the network is forced to execute a simple task, known as pretext task. These tasks aim to extract different feature representations from the images. Therefore, in order to have a good performance, it is important to select an adequate SSL technique depending on the wanted target supervised task. These different tasks will be discussed below.

2.3.5 Geometric Distortion

The distortion technique considers that even if a set of transformations is applied to the image, the image content remains the same [7]. This is important since by applying different transformations to the same image and by forcing the network to predict which one was applied, it is able to extract better visual features from the images. The better the learned features, the better is the knowledge of the network and, therefore, better results the target supervised task will have. The different distortions can be translation, rotation, scaling among others.

For example, if the dataset has a set of deer images, then the network has to identify all the deer images (that are in the same class). To distinguish different classes, the CNN has to learn visual features of each object that contribute to join the images from the same class and distinguish the different classes [7]. Figure 2.3 exemplifies how different transformations applied to an image do not change its content.

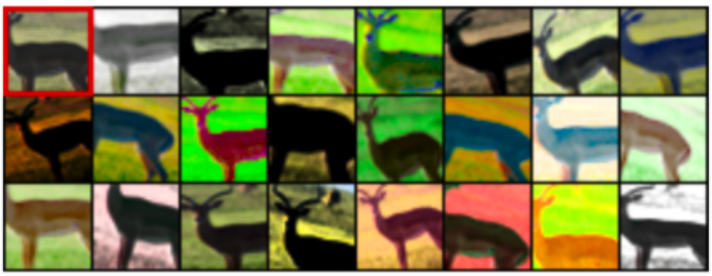


Figure 2.3: The original image is shown in the top left corner, the remaining images are the result of random transformations (extracted from [7]).

Different techniques can be included in these geometric distortion categories. Two examples of works that apply different geometric distortion, the first applies a technique known as Exemplar-CNN [7] and the other uses the Rotation [14] technique.

2.3.6 Patch Relative Position

Another example is the patch technique, which takes advantage of the spatial structure of the image. The key idea behind this technique is that each object, when divided into different parts, should maintain spatial relations between its different positions [25]. This technique considers different patches from the same image and trains a CNN to predict spatial relationships between them.

Different pretexts tasks can be applied. The work described in [8] predicts the relative position between two random patches from the same image. Figure 2.4 exemplifies this technique [8]. Another example [51] is instead of only two patches, it trains the model to predict the relative position of all 9 disordered patches, this is known as the jigsaw puzzle technique. Another pretext task [52] considers visual features as a scalar-value. This technique defined that there was a relationship between counted features in each patch and that different images could be identified by having a distinct number of counted features.

In order to accomplish all the desired pretext tasks, the model is forced to learn features associated with the spatial structure of the image, as well as understand the relationship between different parts of the object and its shape [25].

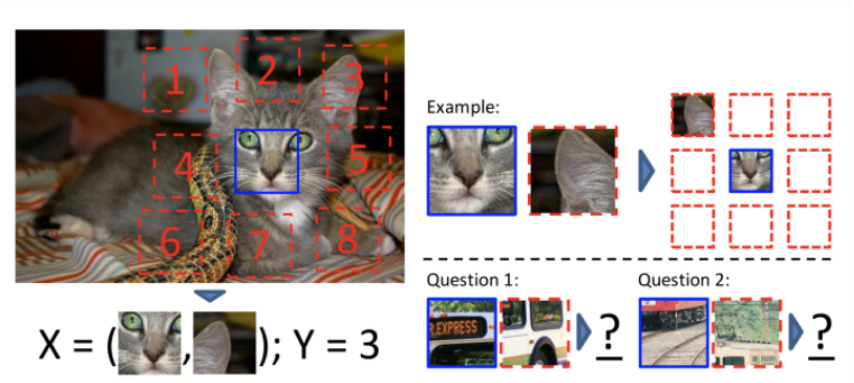


Figure 2.4: Exemplification of the SSL task, on which the network is forced to predict the relative position of two random patches (extracted from [8]).

2.3.7 Colorization

The colorization task emerged as an attempt of extracting features from the images by assigning color to them. In fact, in order to understand the different appropriate colors in an image it is required to detect the different objects [25].

Therefore, it was proposed a simple task, known as colorization [9], this technique given a black and white image asked the model to predict an appropriate color of each pixel. In order to understand which colors are appropriate to each object, the CNN has to understand which object is analyzing by performing object recognition. Figure 2.5 exemplifies the output of the described technique [9]. Another

proposed simple task is the ColorMe technique [24], where the model instead of receiving a black and white image, received as input the green channel of an image and had to predict the red and blue channels colorization of the same image.

Figure 2.5: Example of the output of the colorization technique [9] given a gray scale image. (extracted from [9]).

In the medical world, most of the images are in gray scale, however, dermoscopy images contain color and that is one of the most important aspects in order to identify the malignancy of the lesions. Using colorization as a pre-training task may be adequate for skin image analysis since the network can extract knowledge from the color and texture of the skin [29], which may help the classification task.

2.3.8 Generative Modeling

The main idea of generative modeling is to force the network to reconstruct an image or just part of it while learning feature representations [53]. The main concept is to generate new data from an existing sample distribution and to achieve this the network needs to learn feature representations of the objects.

Generative models include different SSL techniques. The first category uses two different neural networks that are trained to compete with each other, one network generates inputs, the generator, and the other, the discriminator, detects if the input is a real image or if it is an output of the generator. The Generative Adversarial Networks (GANs) [53] and Bidirectional GAN [54] are examples of these techniques. For example, by training these models on human faces, the network is capable of learning variables associated with the human facial expressions [53]. The second category aims to reconstruct an image from a corrupted version of the input, known as denoising autoencoder technique [55], or to inpaint, a missing piece of the original input, known as the context encoder technique [10]. Figure 2.6 exemplifies the result of the context encoder technique [10].

Figure 2.6: Example of the output of the context encoder technique [10], where the network inpaints the missing piece of an image. (extracted from [10]).

The generative modeling techniques, in order to generate new images (similar to the input) and to predict missing parts of an image, reside on the fact that natural images are highly structured [10]. Therefore, this may not be the most adequate method to apply to the skin lesions images, since there are a variety of different textures, colors, and shapes of lesions [20].

2.3.9 Contrastive Learning

Contrastive Learning has proven to be a good technique to extract visual features from the data without human supervision. The main idea of contrastive learning is to focus on high level features rather than paying attention to microscopic details or in other words, pixel-level details. In contrast to the generative methods that measure the loss in the output space, contrastive methods measure the loss in the representation space. This happens since, as the name indicates, contrastive methods rely on contrasting positive and negative samples to learn image representations.

The key intuition idea is similar to the puzzles proposed to children, where it is expected from them to understand the similarity between different versions or views of the same object and detect the dissimilarity of different views of other objects. Figure 2.7 illustrates the intuitive task [15].

Figure 2.7: Intuitive idea behind SimCLR (extracted from [11]).

Contrastive learning uses a score function to measure the similarity between different features, from an image which is similar to the original input image, known as positive sample, and also detect the dissimilarity from an image that is different from the original image, also known as negative sample. There

are many works apply different application of contrastive learning: the Momentum Contrast technique, MoCo, [56], the 'Bootstrap your own latent' technique, BYOL, [57], the CURL technique [58] and the ' A Simple Framework for Contrastive Learning of Visual Representations', SimCLR [59].

All of the discussed techniques could be adequate to apply to the skin cancer diagnoses field. However, the ones that are believed to be more adequate to apply to the skin cancer classification problem are discussed in the section 3.3.

3

Methodology

Contents

This chapter begins with a description of the thesis approach and is followed by the different initialization techniques that were investigated, which will be discussed in further detail in section 3.3. These different initialization methods include two different SSL techniques, which were chosen because they are believed to be more adequate to apply to the skin cancer image analysis.

3.1 Proposed Approach

Most common approaches to the skin cancer diagnosis problem involve supervised learning using CNNs. However, there are not enough labeled images available to train a network from scratch and TL may not be appropriate, since it requires the use of images from different domains. This may lead to a generalization problem caused by the different properties of dermoscopy and natural images.

Figure 3.1 shows an overview of the standard supervised learning pipeline, where the model is trained using a labeled dataset. The architecture consists of an encoder, followed by a FCL to output the various lesion classes. Two strategies can be used to initialize the weights of the encoder: from scratch (random weights) or TL, usually from the ImageNet dataset [27].

Figure 3.1: Thesis proposed framework using only the supervised learning. The dataset from ISIC 2019 [1] [2] [3] will be used.

This thesis proposes to explore a different strategy, which consists in applying SSL, to initialize the weights of the network. Figure 3.2 describes the generic approach for the application of SSL. The first step consists of pre-training a CNN (encoder) using the chosen pretext task and, secondly, fine-tuning the parameters of the pre-trained network to the classification task (this time using labels), by recycling the encoder and adding a FCL to output the various classes available in the ISIC 2019 dataset.

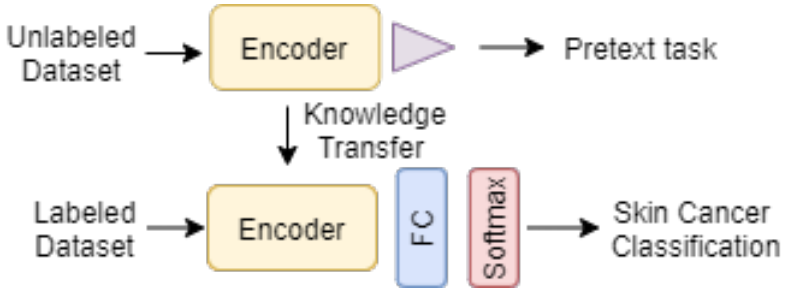


Figure 3.2: Thesis proposed framework using SSL technique applied to the skin cancer diagnosis. The last layers of the pretext task network are replaced by a fully-connected layer to output 8 classes. For the pretext task the unlabeled images from ISIC Archive [12] will be used and for the skin classification task the labeled dataset from ISIC 2019 [1] [2] [3]. The purple triangle represents the last layers of the pretext task architecture.

This thesis will present a comparison between two initialization techniques for the skin cancer diagnosis. Therefore, two techniques will be studied: the Rotation [14] and the SimCLR [59], which are believed to be accurate to apply to the skin cancer problem. A more detailed explanation of each method application, as well as, an image summarizing which part of the CNN will be recycled are presented in 3.3.

In order to confirm the impact that different initialization techniques have on the model performance, a systematic assessment will be performed, both quantitative and qualitative, of six initialization pipelines: two with supervised learning that follow the standard approach found in the literature (random weights and TL from ImageNet) - these will be this thesis baselines - and four self-supervised contenders (two using the SimCLR and other two using the rotation technique, both experiments using random and ImageNet weights as the starting point for the encoder).

There will be also executed two different pipelines that fuse the Rotation and the SimCLR technique, these pipelines are believed to have better results once they combine distinct information from each model. In the section 3.3 will be described the mentioned TL pipelines. The encoder used in these systematic assessments is the ResNet-50.

Apart from these experiments, will also be considered the differences between the learned feature representations among the different training strategies using GradCam [60] and Local Interpretable Model-agnostic Explanations (LIME) [61]. Further details about each algorithm are shown in section 3.4.

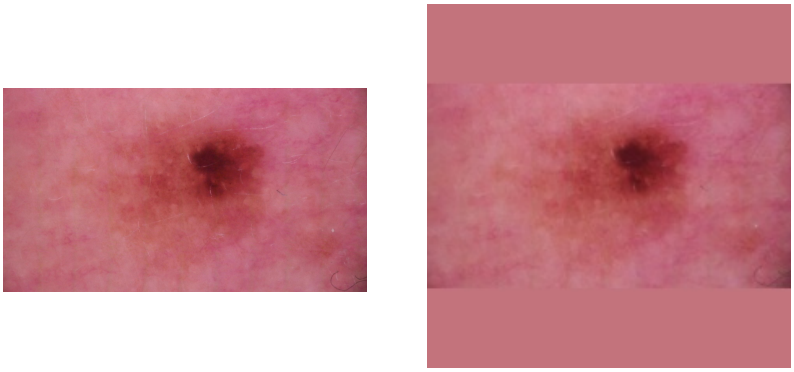
3.2 Data and Training Manipulation

During the execution of this thesis, some issues needed to be corrected both in data and training.

3.2.1 Image Pre-processing

The images presented in the ISIC archive were collected by different medical centers. Since each center generated images with different sizes, colors, and aspect ratios, it was necessary to preprocess the different images. This process normalized the color and allowed all the images to have the same size while maintaining their aspect ratio.

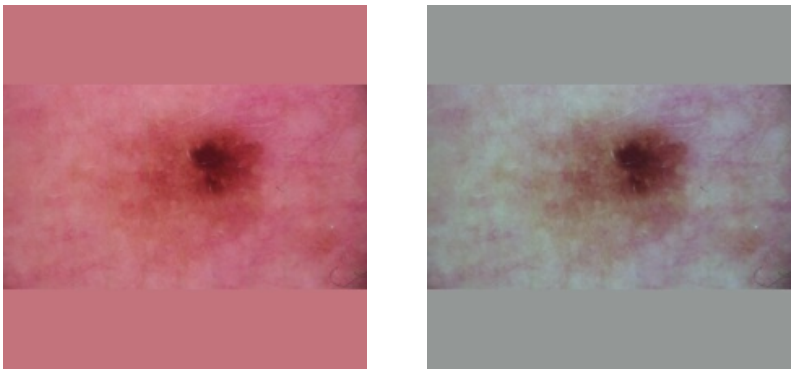
First, all images were converted to squares and, in order to maintain their aspect ratio, there was applied padding of pixels in the smallest margin of the image. The color of the added pixels was selected according to the most predominant color of the image. Secondly, after turning the original image into a square image, then it was resized to a size of 224x224. Figure 3.3 presents an example of the described padding technique.



(a) Original Image (1586x872) **(b)** Padding (224x224)

Figure 3.3: Example of the technique used to convert an image to square while maintaining the initial aspect ratio.

After resizing all the images it was applied the color constancy algorithm Shades of Gray as it is proposed in [13]. The reason why the different images have different colors is mainly due to the light source, therefore this algorithm estimates the color of the illuminant and transforms each image into their canonical light source. Figure 3.4 presents an example of the described padding technique. By applying this algorithm, all the resulting images have similar colors and the same size.



(a) Padding (224x224) **(b)** Shades of Gray (224x224)

Figure 3.4: Example of the color normalization using the Shades of Gray algorithm [13]

3.2.2 Training Specifications

In order to improve both the performance of the supervised and self-supervised classifiers, there has been used the technique of artificially augment the training set which prevents overfitting. This technique creates more variability in the data. To do so random flips (both horizontal and vertical) and rotations of multiples of 90 degrees were performed to all the images presented in the training set. These geometric transformations resulted in an augmentation of the training dataset, which allowed the network to have

better performance.

The used dataset is highly imbalanced, in order to overcome this issue there have been applied class weights to the loss function. This technique assigns to the less frequent classes the higher weight and therefore the loss becomes a weighted average. This allows the model to be more robust since it does not tend to classify all classes with the category that appears more frequently in the dataset. Therefore it promotes a classifier that can learn all classes equally. Equation 3.1 presents the formula used for the class weights.

$$w_i = \frac{N_{total}}{N_i}, \tag{3.1}$$

where w_i is the weight for class i, N_{total} is the total number of samples in the training set and N_i is the number of samples for class i.

3.3 Initialization techniques

This thesis aims to shed a new light on the application of SSL in the skin cancer context. Towards this goal, it was developed a robust experimental framework to:

(i) investigate the impact of SSL on the training and generalization of a CNN for skin lesion diagnosis and demonstrate that even with a small dataset there are benefits in using SSL. In order to better compare the impact of SSL the model was trained using two initialization techniques: i) using random weights and ii) ImageNet weights.

(ii) compare two different SSL approaches, one based on geometric distortion and another on contrastive learning.

(iii) for the first time provide a qualitative assessment of the impact of the different pre-training strategies, using explainability approaches (Gradient-weighted Class Activation Mapping (Grad-CAM) and LIME).

(iv) demonstrate the complementarity of the features learned by the SSL strategies and the benefits of combining them.

Figure 3.5 demonstrates the discussed framework that will be executed in this thesis.

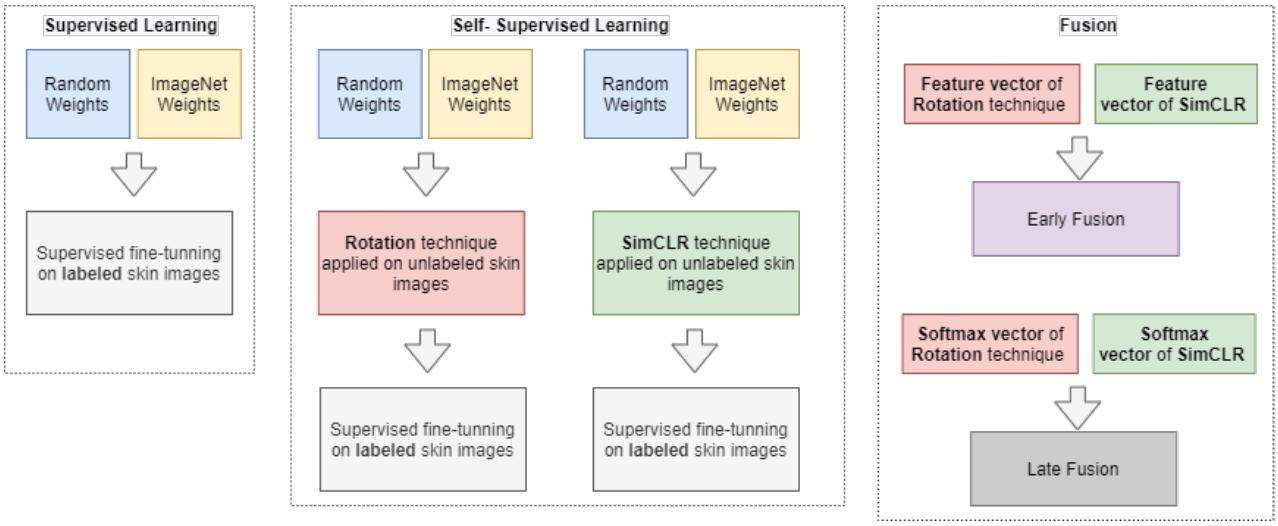


Figure 3.5: Overview of the evaluated pipelines.

This is believed to be the first work to perform a robust quantitative and qualitative validation of the impact of SSL and to demonstrate the importance of combining different SSL techniques. The main idea behind these combination methods is that each model carries different information about distinct aspects of an object and, therefore, combining different models can result in a more robust inference [62]. In other words, when combining a set of models with complementary learned information their performance can have an appealing improvement. In order to combine both Rotation and SimCLR techniques, there will be used early and late fusion.

In the sequence, it will be discussed the different initialization techniques that will be applied in this thesis.

3.3.1 Geometric Distortion

As mentioned before in section 2.3.4, geometric distortion is a SSL technique that takes advantage of a simple set of transformations, which do not change the semantic content of the image. To be able to identify which geometric distortion was applied, the network has to detect characteristics from each object, forcing the CNN to learn semantic features of the image. Therefore, by applying this set of transformations the model is able to extract useful information from the image.

The geometric distortion may have some limitations when applied to the skin cancer images, since these images tend to be less stratified than, for example, a chest x-ray image. Currently, it is being applied rotation techniques to the analysis of thorax images (recall table A.1), where there exists a clear consistency in the structure, (e.g., the heart is always on the left side) and, therefore, the network is able to learn visual features that characterize the structure of the thorax by predicting which rotation was applied [29].

Although the images of skin lesions are less structured, it was opted to apply **the Rotation technique [14]** to the skin cancer classification problem. First due to its simplicity, which is interesting to compare simpler techniques to more complex ones and, secondly, for comparison reasons since it was applied in the previous work [23]. The question that now arises is: How does the self-supervised rotation technique work?

Rotation [14] is a technique that can be seen as a 4-class classification problem, where the network is forced to predict which rotation [0º, 90º, 180º or 270º] has been applied to the image. By learning to distinguish which rotation was applied to the image, the model is forced to identify different details from the input and therefore it extracts useful visual features from the different pictures. Basically, the main intuition for the rotation technique is that, in order to correctly recognize which rotation was applied to the image, the model has to learn to localize and detect the type of object as well as recognize their orientation in the image. By doing so, the model will be able to relate the rotated object with its dominant orientation. Figure 3.6 illustrates the pipeline of the Rotation technique.

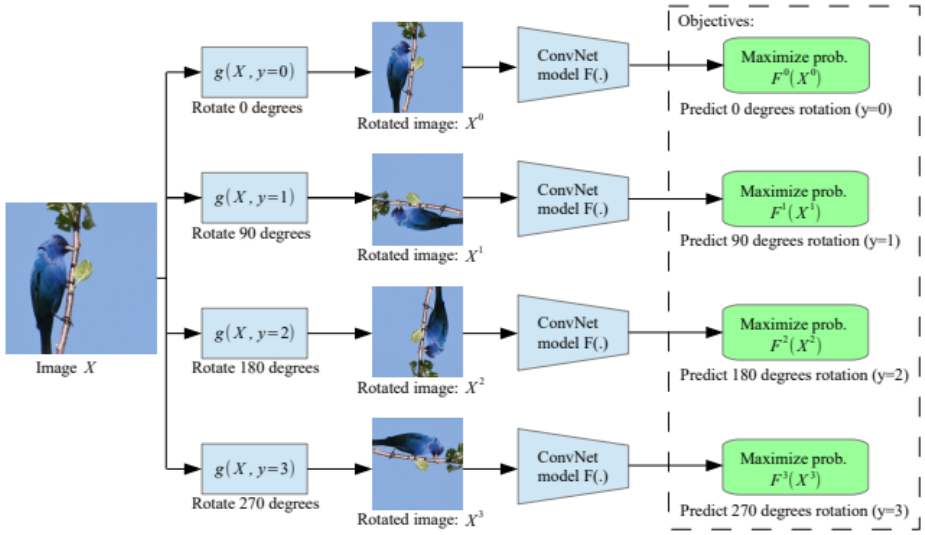


Figure 3.6: Rotation pipeline. The model is represented by f(.) and $f^y(x^y)$ is the probability of the input image being rotated by the y rotation and predicted by model f(.) (extracted from [14]).

Given an input image the goal is to train a model $f(.)$ and force it to estimate which rotation was applied to the original image. Therefore, given a set of N images, the main objective is to find the minimum value of equation (3.2).

$$min_\theta \frac{1}{N} \sum_{i=1}^{N} loss(x_i, \theta), \tag{3.2}$$

where x_i is the input image, θ are the learnable parameters of model $f(.)$ and the loss function, $loss(.)$, is given by equation (3.3).

$$loss(x_i, \theta) = -\frac{1}{K}\sum_{y=1}^{K} log(f^y(g(x_i|y)|\theta)), \tag{3.3}$$

where $K = 4$ is the number of discrete geometric transformations applied to the image, $g(.|y)$ is the term that applies the geometric transformation with label y to the image x and $f^y(.|\theta)$ is the output of the model that gives the probability distribution over all possible geometric transformations.

In this thesis, the encoder will have the ResNet-50 structure followed by a FCL to output the 4 pretended classes that correspond to each rotation applied to the image. Figure 3.7 describes this thesis proposed framework approach for the integration of the self-supervised rotation technique into the skin cancer classification pipeline.

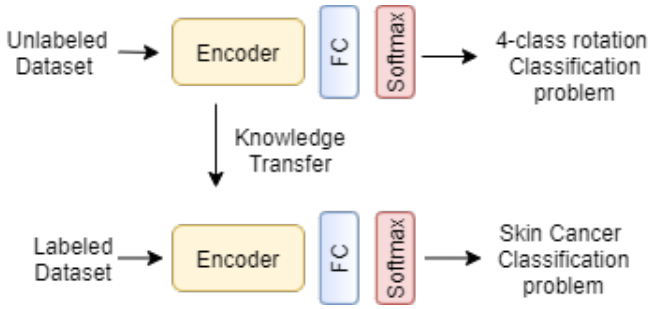


Figure 3.7: Thesis proposed framework using the Rotation technique applied to the skin cancer diagnoses. The last layers of the pretext task network are replaced by a FCL to output 8 classes. For both tasks it will be used the dataset from ISIC 2019 [1] [2] [3], but for the pretext task the labels will not be used.

3.3.2 Contrastive Learning

Contrastive learning is believed to be adequate for the skin analysis application since it focus on high level features rather than focusing on pixel-level details. This way, the network can have a better understanding of each lesion as a whole. In addition, as contrastive learning relies on contrasting positive and negative samples to learn image representations, it is expected to extract visual features from the images that help the network to better discriminate the different classes in the skin lesion classification. Looking at the different works, included in the contrastive modeling, the proposed technique is **the SimCLR [59]** due to its good performance and simple architecture.

Feature representations, in the SimCLR technique, are learned by maximizing the feature agreement between differently augmented views of the same image via a contrastive loss, which will also accentuate the dissimilarity among different images. The key idea is when comparing the multiple images using the contrastive objective, the representations of corresponding views are 'attracted' to one another and the others are 'repelled'.

SimCLR can be divided into four main steps:

1. Sample a mini-batch of n samples, on each batch an image is given as an input, known as x.

Then, random transformations are applied (random cropping, resize the image to its original size with random flip, random color distortion and random Gaussian blur) in order to obtain a pair of two augmented images, x_i and x_j. This pair is considered a positive pair. In the end there will exist $2n$ augmented samples.

2. For each positive pair, x_i and x_j, the remaining $2(n-1)$ images will be used as negative samples. Each augmented image within the pair is sent to an encoder, $f(.)$, in order to obtain the corresponding representations, h_i and h_j, which are the output of the average pooling layer.

3. The obtained representations are then applied to a Multi-layer perceptron (MLP) denoted in [59] as projection head, $g(.)$, to apply transformations and project them into the new space, z_i and z_j. Thus, for each augmented image in the batch there is one vector, z_i and z_j. This space is where the contrastive loss will be applied.

4. The contrastive loss function uses the cosine similarity, given by the following expression:

$$sim(z_i, z_j) = \frac{{z_i}^T z_j}{||z_i||||z_j||}, \tag{3.4}$$

where $||z||$ is the l2 norm of the vector.

Equation 3.4 denotes the pairwise cosine similarity between augmented images. Images with higher similarity will have Cosine similarity values close to 1. The SimCLR uses the so called NT-Xent loss, which stands for normalized Temperature-Scaled Cross-Entropy Loss. This loss function for a positive pair of examples is given by equation (3.5).

$$l_{i,j} = -log\frac{exp(\frac{sim(z_i,z_j)}{\tau})}{\sum_{k=1}^{2N} 1_{[k\neq i]} exp(\frac{sim(z_i,z_k)}{\tau})}, \tag{3.5}$$

where $1_{[k\neq i]}$ is an indicator function that checks if $k = i$ then is 0 otherwise is 1 and τ is a temperature parameter. Lastly, a loss over all pairs is given by equation (3.6).

$$L = \frac{1}{2N}\sum_{k=1}^{N}[l_{i,j}(2k-1, 2k) + l_{i,j}(2k, 2k-1)] \tag{3.6}$$

Based on this loss the representations, h and z, improve over time by approximating the features of similar images in the space.

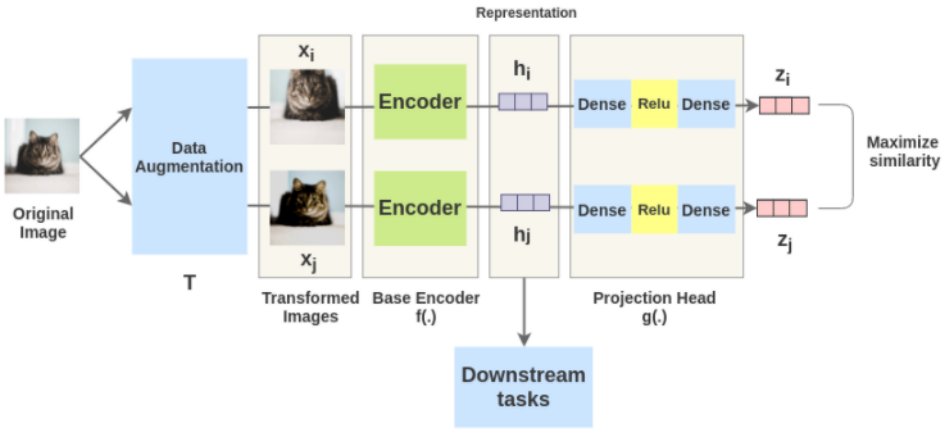


Figure 3.8: SimCLR architecture represented for a positive pair and batch equal to 1 (extracted from [15])

Figure 3.8 contains an arrow pointing to the downstream tasks. Here, downstream tasks correspond to the supervised target tasks to which the self-supervised model may be applied to. Here, it is proposed to apply this SimCLR technique as an initialization of the convolutional layers of the network used for image classification.

In this thesis will be used the ResNet-50 [6] network as the encoder. This network will have to be modified in order to execute this pretext task (e.g., there will have to be added a MLP to project the transformations to the new space z). In the self-supervised pre-training phase, each encoder (with the same architecture), $f(.)$, will receive two augmented images, denoted as x_i and x_j, of the unlabeled ISIC 2019 dataset [63]. Figure 3.9 describes this thesis proposed framework approach for the implementation of the SimCLR technique applied to the skin cancer classification.

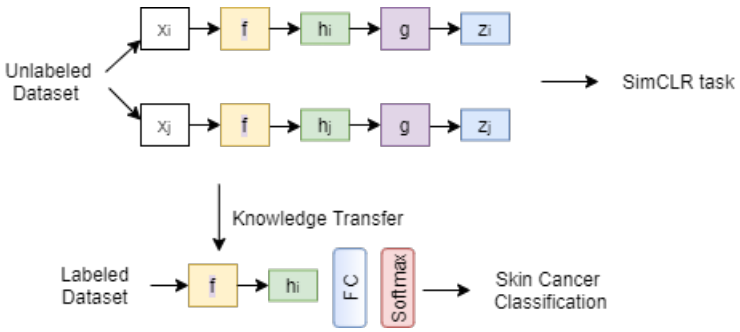


Figure 3.9: Thesis proposed framework using the SimCLR technique applied to the skin cancer diagnoses. The last layers of the pretext task network are replaced by a FCL to output 8 classes. For the pretext task the unlabeled images from ISIC 2019 will be used and for the skin classification task the labeled dataset from ISIC 2019 [1] [2] [3].

3.3.3 Fusion: Rotation and SimCLR

As mentioned before in the beginning of section 3.3, there will be also executed a set of experiments that will combine both SSL techniques: Rotation and SimCLR. It is important to highlight that is also one of the novalties presented in this thesis.

Both SSL techniques force the network to learn different tasks, which results in two models that might learn different information. However, the question that arises is: 'Is the information of both techniques complementary?'.

The goal of conducting these tests is to improve the global performance of both methods, assuming that each model carries different information about each skin lesion. In other words, the combination of both models can result in a more robust inference. To combine both Rotation and SimCLR techniques will be used the early and late fusion approaches, which will be addressed below. Both techniques differ at the level of fusion, early fusion concatenates the models in a feature level, while late fusion fuses the models in the classification scores levels [62].

3.3.3.A Early Fusion

As the name indicates, early fusion combines the different methods in an earlier stage, which is in the feature space. This is known as feature level fusion and it consists of combining all the feature vectors into a single feature vector before sending them to the classifier. Therefore, this new model is trained to learn a correlation between the different features from the input models. In order to combine the set of models, the concatenation was used to jointly represent the different features. Figure 3.10 contains the pipeline used for the combination of the Rotation and the SimCLR technique using the early fusion.

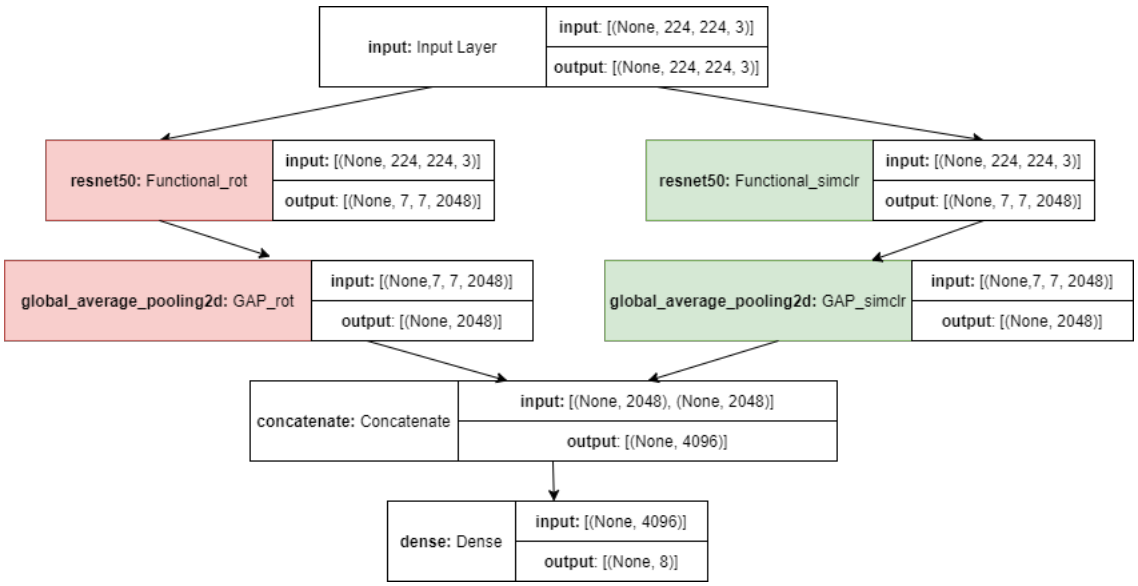


Figure 3.10: Early Fusion Pipeline in red it is represented the rotation model and in green the SimCLR.

3.3.3.B Late Fusion

Late fusion fuses the models in their final level, which is the classification scores level. Usually, this method uses fusion mechanisms that can consist of averaging, voting (which may result only with more than two models), or learned models.

In this thesis, it was decided to apply the mean value to the two score vectors and select the highest value that resulted in the combination of the final decision of both models. Figure 3.11 contains the pipeline used for the combination of the Rotation and the SimCLR technique using the late fusion.

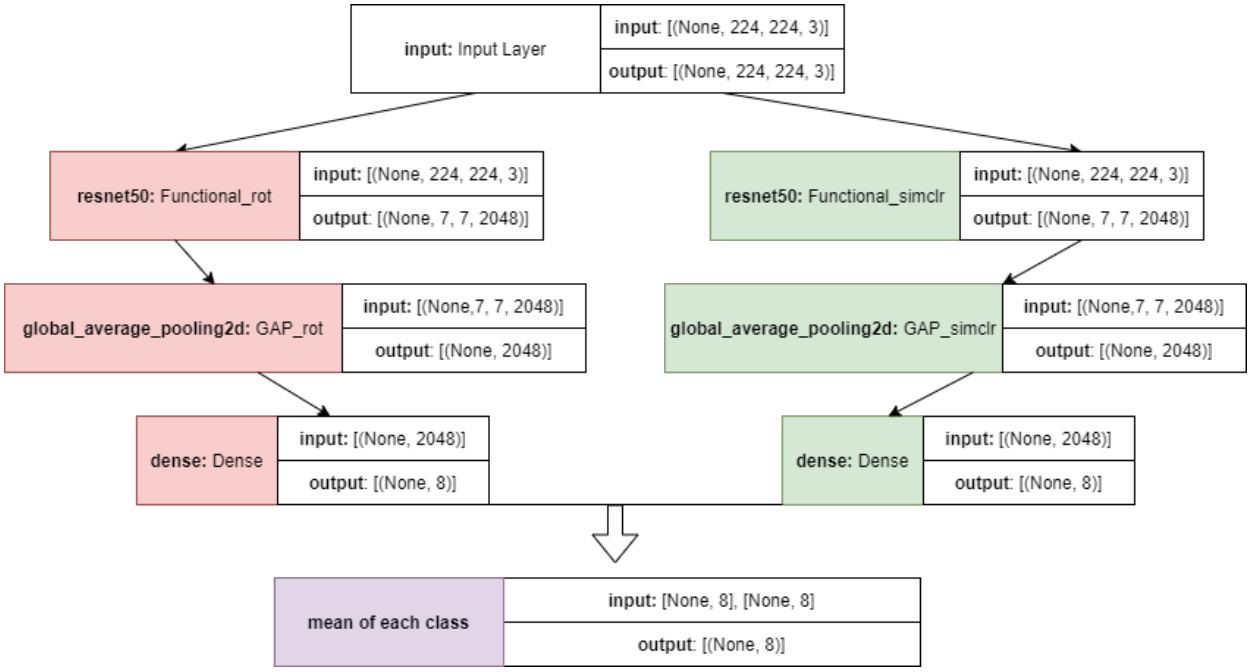


Figure 3.11: Late Fusion pipeline in red it is represented the rotation model, in green the SimCLR and in purple the final score vector that results of mean of both score vectors.

3.4 Feature Assessment

This section will explain two algorithms used to understand what a CNN sees to make a decision (both methods will be used in the qualitative assessment).

3.4.1 Grad-CAM

Grad-CAM [60] uses the gradient information that the last convolutional layers of the CNN have, to determine the importance weights that each neuron has for the predicted class. Therefore, the main goal of Grad-CAM is to explore the spatial information preserved in the convolutional layers to better comprehend the parts of the input that contributed to the predicted decision.

This method could explain activations in any layer of a deep network. However, it is mainly used in the last convolutional layers of the network since these layers have the best compromise between spatial information and high-level semantics.

The output of the Grad-CAM consists of a class-discriminative localization map, $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$, where u is the width and v the height for any class c. Grad-CAM can be divided into three steps:

1. Computing the gradient of the score for class c (before the softmax), y^c, with respect to the feature map activations, A^k, of a convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k}$. Meaning that for a 2D input image, the gradient is 3D, with the same shape as the feature map. There are k feature maps each of height v and width u, therefore the feature maps have shape $[k, v, u]$ which will be the same shape as the gradients.

2. The previous gradients flowing back are global-average-pooled in order to compute the neuron importance weights, α^c_k:

$$\alpha^c_k = \frac{1}{uv} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{i\,j}} \tag{3.7}$$

   α^c_k represents the importance of the feature map k for a target class c. It is important to recall that the gradients have shape $[k, v, u]$ and after doing the pooling, over the height and width, α with dimension k is obtained.

3. Each α^c_k value is used as the weight of the corresponding feature map. The final Grad-CAM heatmap is calculated by doing a weighted sum of the feature maps. A Rectified Linear Unit (ReLU) operation is applied to obtain:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k) \tag{3.8}$$

   Notice that it is applied a ReLU operation, which only considers the positive values of the pixels, since negative pixels tend to belong to other categories in the image.

3.4.2 LIME

LIME [61] is an explanation technique that intents to explain the predictions of a classifier by changing its input and understanding how its predictions are altered.

To ensure that the explanation is interpretable, LIME modifies the original feature space and the interpretable representation. Therefore, $X = \mathbb{R}^p$ is the feature space, $r \in \mathbb{R}^p$ is the original representation of the explained instance and $r' \in X'$ is the interpretable representation. In addition, the model that is being explained is f: $X = \mathbb{R}^p \rightarrow \mathbb{R}$. LIME is able to explain each class separately, hence, in classification, $f(x)$ is the prediction of the relevant class. The explanation model is given by g: $X' \rightarrow \mathbb{R}$ and

$L(f, g, w^r)$ is the loss function that calculates how unfaithful g approximates to f in the locality defined by w^r. The complexity of the explanation $g \in G$ is measured by $\Omega(g)$.

LIME minimizes the loss function $L(f, g, w^r)$, while ensuring that $\Omega(g)$ has a low value to be interpretable by humans. Therefore, the LIME explanation, $\epsilon(r)$, is given by:

$$\epsilon(r) = argmin_{g \in G} L(f, g, w^r) + \Omega(g) \tag{3.9}$$

The LIME approach can be divided into five steps:

1. Generates N samples, which are perturbed samples of the interpretable version of r' (which is the instance being explained).

2. By applying the mapping function is able to recover the previously perturbed observations in the original feature space.

3. Predicts the outcome of every perturbed observation.

4. Computes the weights of every perturbed observation.

5. Solves equation 3.9 using the new dataset, consisting of perturbed samples, with the corresponding response.

As seen in the five steps of LIME, this algorithm generates a new dataset containing perturbed samples and the corresponding predictions. In images, perturbing individual pixels do not make much sense, since many pixels contribute to one class. Therefore, LIME creates variations in the images by first dividing the image into groups of pixels, known as 'super-pixels', and switches them on and off. Super-pixels are interconnected pixels that have similar textures and can be turned off by replacing each pixel with a gray color. Therefore, in images, the interpretable space is a binary vector indicating the presence or absence of a super-pixel. This means that to obtain the explanation of the prediction, the image is perturbed by hiding one or more super-pixels to get the corresponding prediction. Figure 3.12 describes this process.

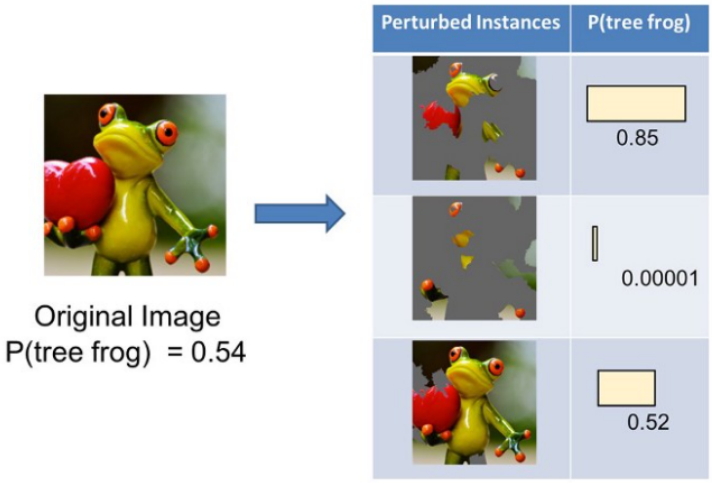


Figure 3.12: Examples of perturbed instances of an image and their predictions (extracted from [16]).

4

Experimental Results

Contents

This chapter starts with a description of the dataset, which will be used in this thesis, it also addresses the evaluation metrics and it is followed by a description and discussion of the experimental results performed during this thesis.

4.1 Dataset

The dataset used during this thesis implementation will be the ISIC 2019 [1] [2] [3] dataset. However, for the pre-training implementation using self-supervised techniques, it will be used the same dataset but without the labels. ISIC 2019 is a set of dermoscopic images of skin lesions, with medically annotated labels. In particular, each image was classified into one of 8 classes already discriminated in section 1.2. Figure 4.1 illustrates eight different examples of skin lesion for each of the 8 different lesions classes: Actinic Keratosis (AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevus (NV), Squamous Cell Carcinoma (SCC) and Vascular (VASC). The dataset also contains patient's metadata, however, this information will not be considered during the implementation of this thesis.

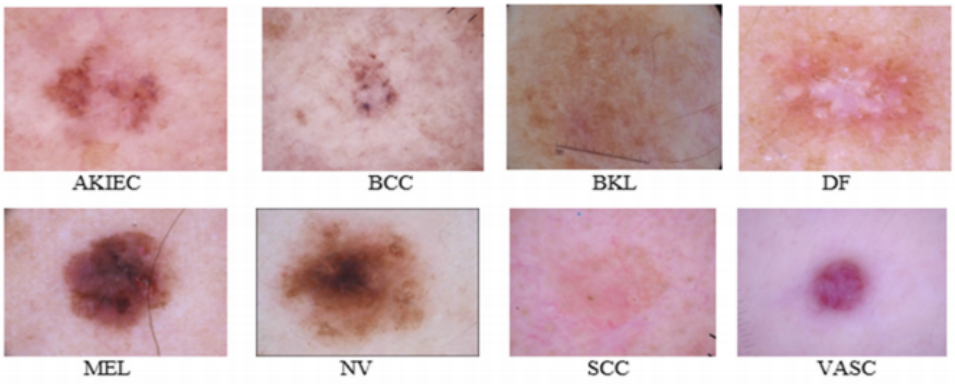


Figure 4.1: Examples of skin lesions from 8 different classes (extracted from [17]).

The dataset contains a total of 25,331 images containing ground truth labels for training and a total of 8,238 images for testing. It is important to note that the training set images contain medically annotated labels and the testing set does not provide labels. Table 4.1 contains information about the training and testing set.

Table 4.1: Total Number of Samples in the Training and Test sets.

Dataset	Total	MEL	NV	BKL	DF	VASC	BCC	AKIEC	SCC
Train	25,331	4,522	12,875	2,624	2,39	253	3,323	867	628
Test	8,238								

4.2 Evaluation Metrics

The process of overfitting happens when a model is built with features that have high accuracy in the training data but very little performance when used with new data. Cross-validation is a strategy that tends to mitigate the occurrence of overfitting. There are different cross-validation strategies, for example, the K-fold approach uses only each data once in the validation dataset, in other words, there are no repetitions. Another example is the Monte Carlo sampling which contains repetitions in the validation set. During this method, each split is independent from the other in a way that the original data is split randomly for each different fold, even if it contains some repeated images.

To compare the different initialization techniques it was opted to use the Monte Carlo sampling technique since it was desired to evaluate the different methods and there was no problem with repeating images between sets. It was opted to use 5-partitions since each method takes a considerable amount of time to run. Each fold was created by dividing the original training set into a smaller training set (70%) and a validation set (30%).

In figure 4.2 it is possible to see the division of the training and the validation dataset for one partition. All five partitions have been confirmed to be different and they all have the same number of images for each class.

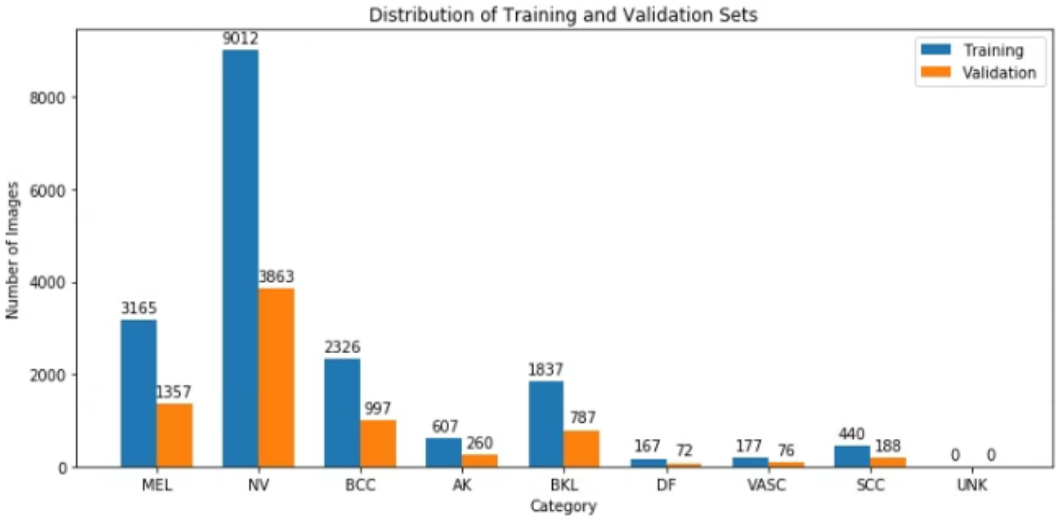


Figure 4.2: Example of the distribution of the training and validation sets for each partition.

Below it will be explained the different evaluation metrics: Balanced Accuracy (BACC), Precision, F1-Score, SP, and Area Under the Curve (AUC).

4.2.1 Confusion Matrix

To compute the performance metrics the confusion matrix was performed, this has dimension $k \times k$, where k is the number of classes. To compute this matrix, there has to exist a one-vs-all strategy for each class. Each matrix entrance, ij, contains the probability of predicting class, j, when the real class is i. The one-vs-all strategy consists in assuming the positive class vs all the remaining classes. Therefore, while making these comparisons four parameters can be taken from the confusion matrix: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) (from the point of view of each class). Figure 4.3 represents an example of a 8-class problem.

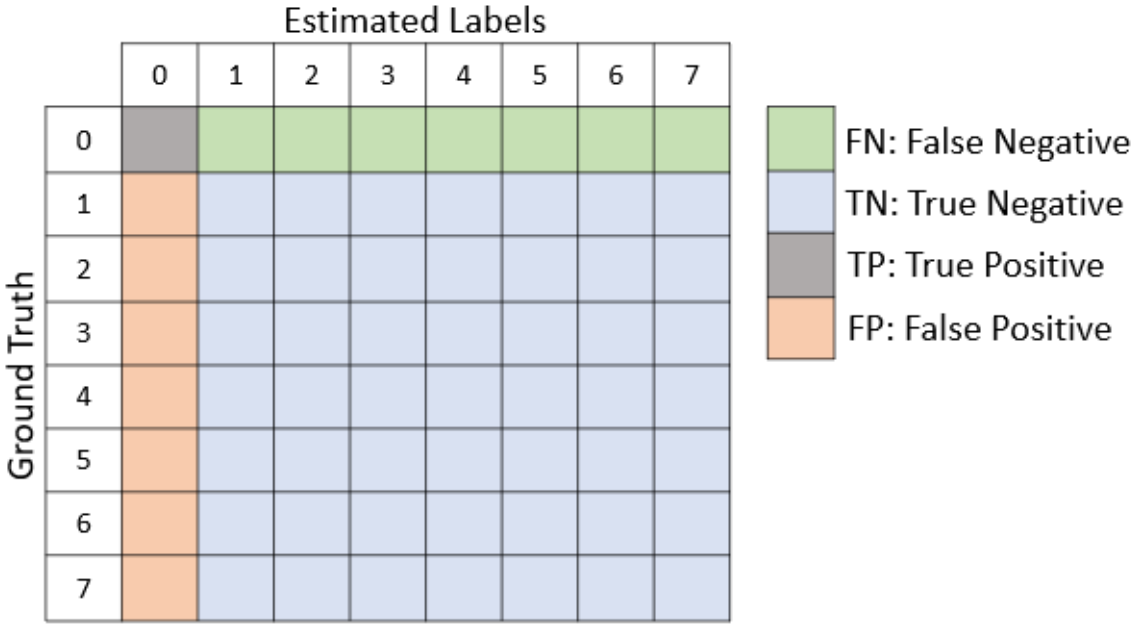


Figure 4.3: Example of a 8-class confusion matrix, where the positive class is class 0. Each component corresponds to the sum of the same color cells.

4.2.2 BACC

The used dataset is unbalanced, therefore, instead of using the weighted accuracy, the BACC was used. This metric gives equal importance to all classes independently of the available number of examples. BACC averages the Sensibility (SE) obtained for each class, SE_i (also known as *recall*). The SE_i is given by:

$$SE_i = \frac{TP_i}{TP_i + FN_i}, \tag{4.1}$$

where TP_i is the number of true positives, i.e, the number of correctly classified examples from class i and FN_i is the false negatives for class i.

Therefore, BACC is given by:

$$BACC = \frac{\sum_{i=0}^{7} SE_i}{8}. \tag{4.2}$$

4.2.3 Precision

The precision measures which fraction of all the positive predicted records is actually positive. This metric is given by equation 4.3.

$$precision = \sum_{i=0}^{7} \frac{TP_i}{TP_i + FP_i}, \tag{4.3}$$

where TP_i are the true positives and FP_i are the false positives for class i.

4.2.4 F1-Score

F1-score can be seen as a weighted average of the precision and recall, where it reaches its best value at 1 and worst score at 0. F1-score is given by:

$$F1 - score = \frac{2 * (precision * recall)}{(precision + recall)} \tag{4.4}$$

4.2.4.A SP

SP measures the true negative rate and it gives the negative samples that were correctly classified. SP for each class is given by:

$$SP_i = \frac{TN_i}{TN_i + FP_i} \tag{4.5}$$

The final SP value is obtained:

$$SP = \frac{\sum_{i=0}^{7}(SP_i)}{8} \tag{4.6}$$

4.2.5 AUC

AUC score measures the ability of a model to distinguish between classes and is used as a summary of the ROC curve [64]. ROC stands for "Receiver Operator Characteristic" and it gives the balance between the true positive and the false-positive rate of a classifier. A model with good performance yields a high AUC score and it means that the model is capable of distinguishing between the positive and negative classes.

4.3 Computational Environment

The experiments were carried out using the programming python language, using the libraries: Keras [65] and Tensorflow [66]. The Google Colaboratory tool [67], that is a python notebook from Google Research was used in most experiments. This is a platform that provides free access to a NVIDIA Tesla K80 GPU. However, for the complementary study with the use of more data, it was opted to use a laptop computer with the following specifications: Processor: AMD Ryzen Threadripper 3960X 24-core; Memory: 128 GB RAM; Graphics Processing Unit (GPU): NVIDIA GeForce RTX 3090.

4.4 Effect of image transformations on SimCLR technique

The authors of SimCLR [59] stated that with the composition of data augmentation, the contrastive prediction task becomes harder to execute, however, there is an improvement of the quality of the representations.

Therefore, to first understand the effects of individual data augmentations in the SimCLR technique and, secondly, to understand the importance of combining those data augmentations, preliminary experiments were performed. It was opted to use a SVM since the objective is to evaluate the SimCLR pretext task only, in other words, the feature space is being evaluated.

First, it was applied to the input image horizontal flips, then the composition of those flips with central crops and finally it was added rotations of 0, 90, 180, or 270 degrees. It was also studied the impacts of random color distribution and random gaussian blur, however, these experiments resulted in a lower performance of the model. This could be explained by the fact that in the skin cancer classification problem the color of each lesion, as well as their sharpness, is highly important. Figure 4.4 illustrates the implemented augmentation compositions.

Apart from the discussed experiments, it was also evaluated the performance of the network in the different feature spaces. These spaces were described in section 3.3.2: first, the 2048 vector, corresponds to h_i and h_j vectors; the 256 feature space to the output of the first dense layer of the projection head, $g(.)$, and finally the 128 space, is described as the z_i and z_j vectors. As the original paper stated, it was possible to confirm that the feature space which had better performance was the 2048 feature space, and therefore the classification task was applied in the space formed by the h_i and h_j vectors.

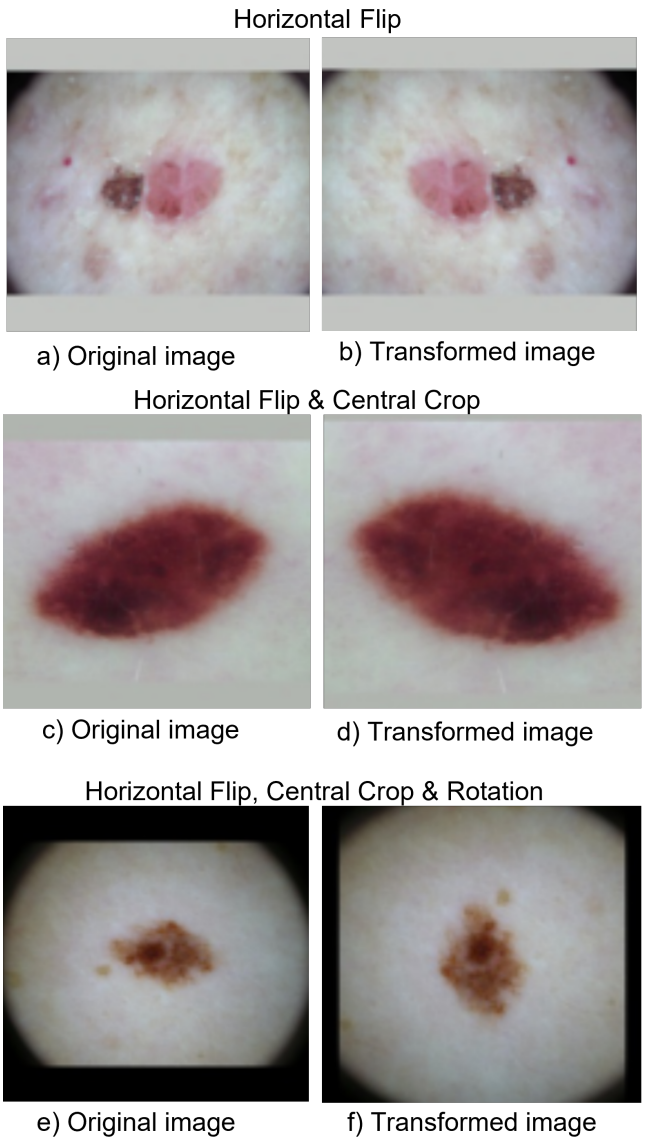


Figure 4.4: Example of the implemented data augmentation while using the SimCLR technique.

Tables 4.2, 4.3 and 4.4 contain the set of experiments that were executed in order to verify the importance of combining different data augmentations. These data augmentations were evaluated using a SVM in a 2048 dimensional space and were executed with a Gaussian RBF kernel with different combinations of two parameters: gamma and C. Gamma is the parameter that decides how much curvature the decision boundary has, the highest gamma the more curvature it has. The C parameter determines the amount of data samples that are allowed to be placed in different classes. In other words, it controls the flexibility of having data points on the wrong side of the boundary. A low value of C maintains a smooth classification, on the other hand, a high C tries to minimize the misclassification of training data.

Analyzing tables 4.2, 4.3 and 4.4, it is possible to see that the highest BACC, for all three cases, is with parameter gamma = 0.1 and C = 1, which is highlighted in bolt. As mentioned in paper [59], it was possible to confirm, from visualizing all three tables, that with the combination of data augmentations the accuracy tends to improve and that is why the following experiments were executed using SimCLR with horizontal flips, central crops and rotation of multiple of 90 degrees.

Table 4.2: Evaluation of the 2048 dimensional space of the SimCLR pretext task using the SVM with Gaussian RBF kernel and the data augmentation: **horizontal flips**.

Parameter		train		validation	
g	**c**	**ACC**	**Balanced ACC**	**ACC**	**Balanced ACC**
0,01	1	0,3386	0,3906	0,3089	0,2602
	50	0,6735	0,8268	0,4278	0,2314
	100	0,7449	0,8806	0,4280	0,2212
0,1	1	0,6385	0,7764	0,4649	**0,2432**
	50	0,9999	0,9999	0,5309	0,2092
	100	0,9999	0,9999	0,5308	0,2091
1	1	0,9999	0,9999	0,5103	0,1282
	50	0,9999	0,9999	0,5104	0,1283
	100	0,9999	0,9999	0,5104	0,1283

Table 4.3: Evaluation of the 2048 dimensional space of the SimCLR pretext task using the SVM with Gaussian RBF kernel and the combination of data augmentation: **horizontal flips and central crops**.

Parameter		train		validation	
g	**c**	**ACC**	**Balanced ACC**	**ACC**	**Balanced ACC**
0,01	1	0,3685	0,3886	0,3437	0,2282
	50	0,6676	0,8200	0,4303	0,2341
	100	0,7061	0,8509	0,4305	0,2303
0,1	1	0,5877	0,7393	0,4446	**0,2542**
	50	0,9955	0,9984	0,5321	0,2256
	100	0,9975	0,9991	0,5296	0,2231
1	1	0,8392	0,8646	0,4995	0,2091
	50	0,9999	0,9999	0,5317	0,1558
	100	0,9999	0,9999	0,5317	0,1558

Table 4.4: Evaluation of the 2048 dimensional space of the SimCLR pretext task using the SVM with Gaussian RBF kernel and the combination of data augmentation: **horizontal flips, central crops and rotation.**

Parameter		train		validation	
g	**c**	**ACC**	**Balanced ACC**	**ACC**	**Balanced ACC**
0,01	1	0,3355	0,3667	0,3163	0,2737
	50	0,6662	0,8143	0,4518	0,2629
	100	0,7045	0,8456	0,4551	0,2561
0,1	1	0,5655	0,7165	0,4489	**0,2844**
	50	0,9989	0,9997	0,5392	0,2414
	100	0,9999	0,9999	0,5400	0,2407
1	1	0,8708	0,9559	0,5161	0,2398
	50	0,9999	0,9999	0,5680	0,1983
	100	0,9999	0,9999	0,5680	0,1983

4.5 Training Conditions

During this work, studies were performed to compare different initialization techniques. In order to execute these studies, there were implemented two different tasks: the pretext task (Rotation and SimCLR technique) and the supervised classification task. The common conditions used for both tasks will be addressed below.

4.5.1 Unsupervised Pretext Tasks

In the different used pretext tasks the following common conditions were implemented:

- The loss used for the rotation technique was the sparse categorical cross-entropy and for the SimCLR was the NT-Xent loss and both techniques used the Adaptive Moment Estimation (Adam) Optimizer algorithm.
- The batch size is equal to 32.
- The training for the rotation technique was performed during 40 epochs and the SimCLR was trained until the loss of the validation set stopped improving.
- A dropout layer with p = 0.5 is used.

4.5.2 Supervised Skin Lesion Classification task

During the supervised classification task the following common conditions were implemented:

- The loss is the categorical cross-entropy with the Adam Optimizer algorithm.
- The batch size is equal to 32.
- The training was performed during 60 epochs
- The class weights as described in equation 3.1 are used.
- A dropout layer with p = 0.5 is used before the softmax layer.

4.6 Comparison between the different initialization techniques

This section is divided into two parts: i) a quantitative analysis, where a comparison between the different approaches taking into consideration the selected evaluation metrics is made; ii) a qualitative analysis that used the Grad-CAM technique [60] to convey a more interpretable analysis of the impact of the various initialization strategies in the features learned by the model;

4.6.1 Quantitative Analysis

Table 4.5 summarizes the median and standard deviation (across all partitions) of the scores obtained for the different initialization techniques. Analyzing table 4.5 it is possible to elaborate different comparisons that will be addressed below.

Table 4.5: Application of the Monte Carlo Sampling with different initialization techniques: training the model from scratch or fine-tuning with ImageNet weights; application of two SSL techniques -Rotation and SimCLR.

Initialization	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
Baseline	Scratch	46,82 $\pm$ 2,00	35,37 $\pm$ 3,84	37,24 $\pm$ 4,64	92,89 $\pm$ 0,55
	Imagenet	71,48 $\pm$ 1,82	65,14 $\pm$ 2,78	67,93 $\pm$ 1,75	96,04 $\pm$ 0,12
Scratch + SSL	Rotation	54,92 $\pm$ 1,15	40,54 $\pm$ 1,84	43,19 $\pm$ 2,04	93,39 $\pm$ 0,18
	SimCLR	52,54 $\pm$ 0,86	44,62 $\pm$ 1,39	47,53 $\pm$ 0,96	93,94 $\pm$ 0,18
Imagenet + SSL	Rotation	71,47 $\pm$ 0,30	62,37 $\pm$ 0,74	65,70 $\pm$ 0,47	95,77 $\pm$ 0,05
	SimCLR	65,51 $\pm$ 0,55	54,47 $\pm$ 2,71	58,28 $\pm$ 1,95	95,17 $\pm$ 0,18

4.6.1.A Trained from Scratch vs fine-tuned with ImageNet

By looking at table 4.5 it is possible to see that fine-tuning from the ImageNet weights is beneficial in terms of performance and stability. In fact, the results using the ImageNet weights (row 2, 5, 6) tend to have a higher median and lower standard deviation when compared to the ones trained from scratch (row 1, 3, 4).

4.6.1.B Supervised vs Self-supervised training

It is also possible to confirm that there are some benefits while using SSL when compared to the standard training. For this evaluation, it will be executed two comparisons. The first one with the models trained from scratch and the second one with the models pre-trained in ImageNet.

Models trained from scratch:

By looking at the baseline trained from scratch (row 1) and to both rows trained from scratch with SSL techniques (row 3 and 4) it is visible that both self-supervised techniques presented higher median and lower standard deviations. Looking at the BACC, the model pre-trained with the SimCLR technique (row 4) had an increase of approximately 6% and the one with the rotation had an increase of 8%, when both compared to the baseline (row 1). In terms of stability, both were approximately 1% more stable than the baseline. This proves that when comparing models trained from scratch there is a tendency to have higher accuracy and more stability (the standard deviation has a lower value) in the models that use SSL. However, looking at the remaining metrics (Precision, F1-score and SP) it is interesting to verify that the SimCLR technique has a better performance.

Models fine-tuned with ImageNet weights:

By looking at the models trained using the ImageNet weights: the baseline (row 2) and to both models that used the SSL techniques (row 5 and 6) it is visible that the latter two have higher stability (lower standard deviation) for all metrics. Looking at the model that used the SimCLR technique (row 6) and analyzing the BACC, it is visible that it had lower accuracy than the baseline (-6%). This is a complicated model that uses a contrastive loss in the pretext task and, therefore, it could have benefited more from a higher number of unlabeled images. However, when looking at the standard deviation of the SimCLR model it is possible to see that it was approximately 1% more stable than the baseline. Analyzing the model with the rotation technique (row 5) it is possible to see that it had similar BACC to the baseline, however, this model had higher stability (lower standard deviation of 2%), which means that despite being similar it is more beneficial to use the rotation technique. Looking at the remaining metrics (Precision, F1-score, and SP) it is possible to verify, then again, that both models have a lower median, but tend to have more stability when compared to the baseline. It is very beneficial to have a more stable model since this makes the model more trustworthy to apply to other data.

This proves that when comparing models trained with the ImageNet weights there is a tendency to have more stability (the standard deviation has a lower value) in the models that use SSL.

4.6.1.C Rotation vs SimCLR technique:

Looking at the self-supervised models trained from scratch (row 2 and 3) it is possible to see that the rotation technique has a higher BACC (2%), however, it is a bit less stable since the standard deviation has a higher value of about 0.3%. The same can be confirmed by looking at the models first pre-trained using the ImageNet weights and then pre-trained with the self-supervised techniques (rows 5 and 6). In this case, the Rotation continues to have higher accuracy (6%) when compared to the SimCIR model. However, with the ImageNet initialization the Rotation technique is a bit more stable (0.2%). Looking at the remaining metrics (Precision, F1-score, and SP) it is visible that the Rotation technique is also the one with the best performance.

These results show that there are benefits while using SSL since there is less variability in the performance of the classifier. This proved that when combining TL with SSL the generalization problem that occurs when using TL is filtered. As mentioned before, TL uses natural images that have a different domain to the skin lesion ones. Therefore the network resulted from applying TL, will have neurons that remain loyal to the natural images. By applying SSL these neurons are 'corrected' and the obtained network generalizes better to the skin lesion images.

4.6.2 Qualitative Analysis of the learned representations

It was opted to execute a qualitative analysis to understand what each model saw differently and what it learned in order to make a decision. The Grad-CAM algorithm [60] was used to analyze the differences between the representations learned by each technique. This is a technique used for visualizing where a convolutional neural network model is looking. During this analysis, it was opted to choose the layer *'conv5_block3_out'*, since it was the last layer of the ResNet-50 and, therefore, its corresponding heat-map displays the most accurate visual explanation of the object being classified by the model.

4.6.2.A How to analyze the Grad-CAM algorithm?

First, it is important to recall that the output of Grad-CAM is a heat-map visualization for a given class label. The given heat-map highlights the parts of the image that the CNN is looking at. Therefore, this is an important tool, since it allows the user to visually verify the focus of the network. During this visualization process, the VIRIDIS colormap was used. Figure 4.5 shows the selected colormap.

Figure 4.5: For the Grad-CAM heat-map the VIRIDIS color map was used, in order to visualize deep learning activation maps with Keras and TensorFlow. The yellow color corresponds to the higher values and, therefore to a higher activation and the dark blue to smaller values which correspond to a lower activation.

To ease the comprehension of how to analyze the Grad-CAM heat-map it will be presented, in figure 4.6, an example of a skin lesion and its corresponding heat-map obtained using Grad-CAM algorithm. First, it is important to recall that the heat-map given by the Grad-CAM shows both the importance given to different aspects in the image as well as the capability of detecting, in this case, the skin lesion.

(a) Original Image **(b)** Heat-map **(c)** Overlap

Figure 4.6: Example of a Grad-CAM heat-map obtained from the SimCLR pre-trained model (layer_name = conv5_block3_out).

Figure 4.6 (a) represents the original skin lesion; (b) the heat-map obtained in the model and (c) represents the overlap of the heat-map with the original image. In fact, by looking at figure 4.6 (b), it

is visible that the model gives higher importance to the center of the lesion (in yellow), proving that this model is quite accurate in detecting the skin lesion and also, despite being highlighted in dark blue, the model gives lower importance to the margins of the image.

To simplify, the next figures will only present the original image, (a), and the overlap of the Grad-CAM heat-map, (c).

4.6.2.B Visualization and interpretation of the learned representations using Grad-CAM

In order to demonstrate the differences between the learned representations two out of the five partitions were examined. In every figure, there will be presented the different models and their corresponding heat-map for each partition. In this section, it was opted to only analyze the models initialized with ImageNet weights since they have better results.

Figure 4.7 shows the Grad-CAM heat-map obtained from the different models (fine-tuned with ImageNet weights) for four skin lesions[1]. By analyzing this figure it is possible to observe that for the same input image all three models look at different parts of each lesion. Therefore, apart from having different performances, each model learns different information about each class of lesion. The SimCLR pre-trained model tended to focus more on the parts of the lesion that presented higher contrast, while the Rotation (visible in the second row for both partitions) looked more at the structure of each lesion. The ImageNet pre-trained model, was the least intuitive to interpret since its focus varied between lesion and skin.

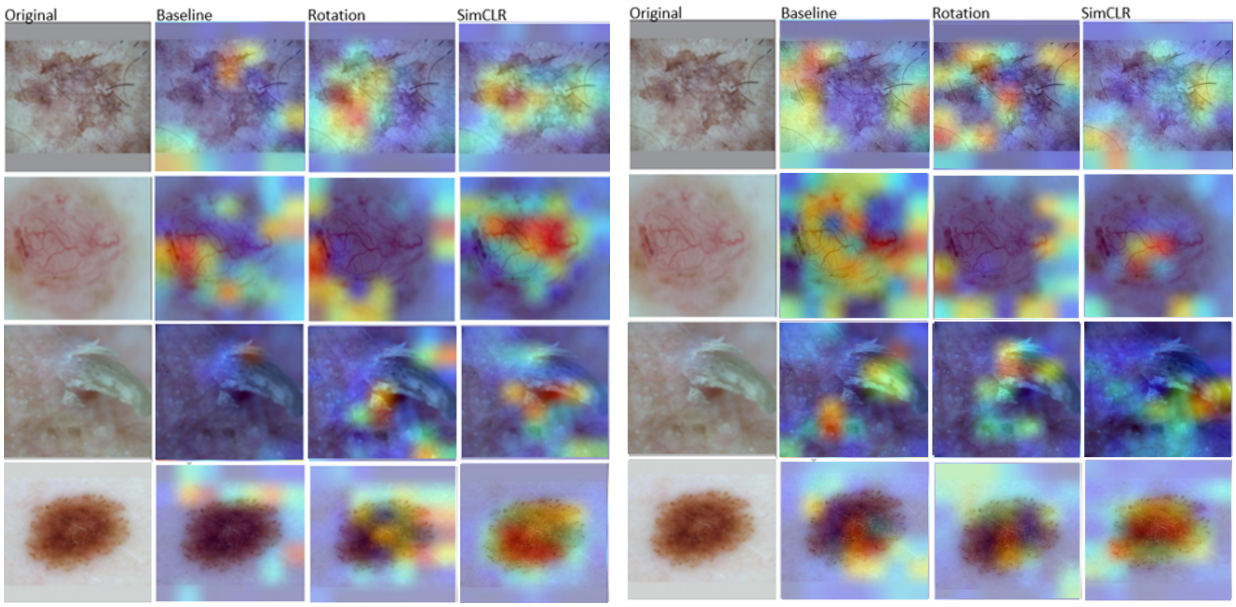
Limitations of each SSL pre-trained model

After, analyzing a set of different images it was visible that each method had some limitations. Figure 4.8 presents some examples that highlight the limitations of each SSL pre-trained model.

It was observed that the model pre-trained with the Rotation technique tended to have difficulties in detecting centered and symmetrical lesions. Each rotation of 90 degrees in symmetric lesions is similar, therefore the model does not learn useful information about these types of images. This limitation is visible in the third and fourth rows of figure 4.8.

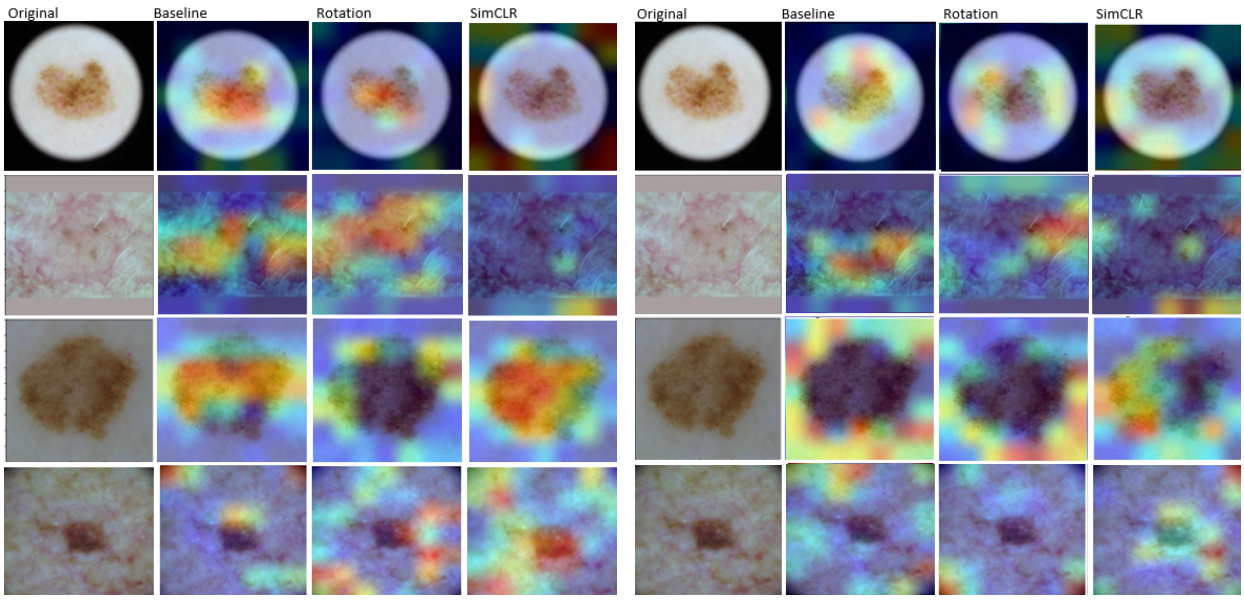
The models pre-trained using the SimCLR technique showed to be more precise in detecting the lesions since the lesions tended to have higher contrast to the skin. However, as some images contained margins with higher contrast this method tended sometimes to focus more on the margins than the lesion itself. This is visible in the first and second row of figure 4.8. Despite being the most precise method, out of the three, in detecting the different skin lesions, the SimCLR method had a lower performance. This means that despite being able to distinguish the lesions from the surrounding skin, this model was unable to learn discriminative information to classify the correct class of lesion.

[1] The models were trained in different partitions, but the shown images are from the same dataset



(a) First Partition **(b)** Second Partition

Figure 4.7: Example of different lesion visualizations using the Grad-CAM algorithm, which localizes class-discriminative regions of each model (Baseline, Rotation and SimCLR).



(a) First Partition **(b)** Second Partition

Figure 4.8: Example of different lesion visualizations using the Grad-CAM algorithm, highlighting the limitations of both SSL methods.

4.7 Fusion of SSL Approaches

As expected, the qualitative assessment proved that apart from having different performances, each network ends up learning different information about the images. Therefore, this led us to the following question: Is the learned information of both techniques complementary?

In order to answer this question, both SSL technique were fused. Therefore as a consequence of this interrogation, there were performed two tests that fused the two models pre-trained with SSL. First, the early fusion was used, which fuses the different methods in the feature space. Secondly, it was used the late fusion that fuses the models in the classification scores level (applied the mean strategy).

The results were evaluated with a quantitative and qualitative analysis. Therefore, this section is divided into two parts: i) a quantitative analysis of the fusion of SSL strategies; ii) a qualitative analysis that used the LIME algorithm [61] to convey a more interpretable analysis of the impact of the different strategies in the features learned by the model.

4.7.1 Quantitative Analysis of the fused models

The results of both fusions appear in the last two rows of table 4.6 (row 3 and 4). It is possible to conclude that, looking at the BACC, the **early fusion had better results than any other model both in stability and accuracy**, proving that, in fact, the features of both models have complementary information. However, the late fusion proved to have worse results, meaning that the features are complementary, but not the learned classification models.

Table 4.6: Application of the Monte Carlo Sampling with different initialization techniques (using ImageNet weights): application of two SSL techniques -Rotation and SimCLR- and fusion of both techniques.

Initialization	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
Imagenet + SSL	Rotation	71,47 ± 0,30	62,37 ± 0,74	65,70 ± 0,47	95,77 ± 0,05
	SimCLR	65,51 ± 0,55	54,47 ± 2,71	58,28 ± 1,95	95,17 ± 0,18
Fusion	Early Fusion	**73,78 ± 0,24**	**68,41** ± 4,13	**70,99** ± 2,61	**96,40** ± 0,36
	Late Fusion (mean)	57,09 ± 2,19	50,28 ± 1,41	52,02 ± 1,08	94,24 ± 0,19

4.7.2 Qualitative Analysis of the fused models

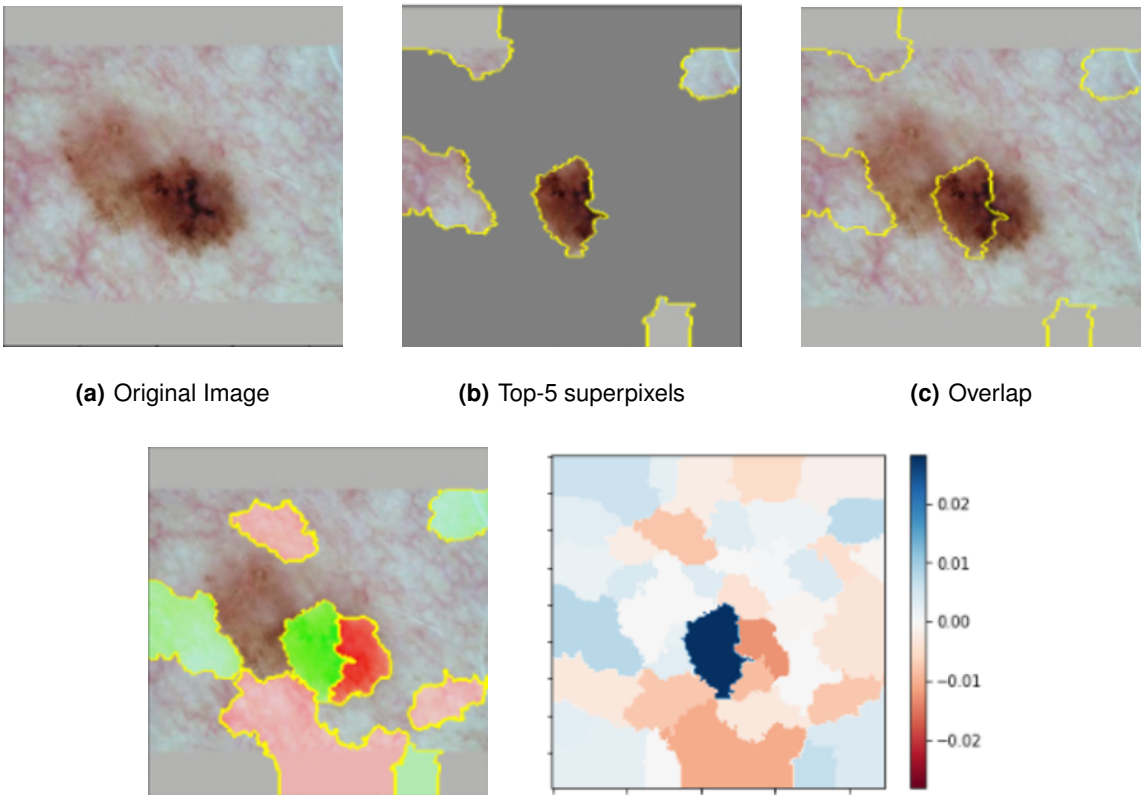
To analyze the learned representations of the fused model, the LIME algorithm [61] was used. It was not possible to use Grad-CAM since for visualization purposes this method used the last layer of the ResNet and the fused model has two ResNet-50. LIME is a model-agnostic, which means that is applicable to any machine learning model. It was opted to only analyze the model that used Early Fusion since it had better results.

4.7.2.A How to analyze the LIME algorithm?

It is important to recall that LIME attempts to interpret a model by changing its input and understanding how its predictions are altered. This model provides local model interpretability. Therefore, it changes the data sample by altering the feature values and it observes the impact that each alteration has on the output. By doing so this method is able to detect which features are important. During this visualization process, the RdBu diverging colormap was used. Figure 4.9 shows the selected colormap.

Figure 4.9: For the LIME heat-map the RdBu color map was used. The blue color corresponds to the higher values and the red to smaller values.

To ease the comprehension of how to analyze LIME it will be presented, in figure 4.10, an example of a skin lesion and its corresponding LIME output.



(a) Original Image **(b)** Top-5 superpixels **(c)** Overlap

(d) Pros in green and Cons in red **(e)** Heatmap containing the explanation weights

Figure 4.10: Explaining an image classification made by the prediction of the SimCLR pre-trained model. The top class was AKIEC.

Figure 4.10 (a) represents the original skin lesion; (b) the top-5 super-pixels that are most positive towards the predicted class with the rest of the image hidden; (c) represents the overlap of the top-5 super-pixels with the original image; (d) represents the pros (green) and cons (red) with weight at least 0.1 and (e) which plots the explanation weights onto a heat-map visualization. In fact, by looking at figure 4.10 (b), it is visible that the model detects the center of the lesion, which correspond to the pixels with a higher importance in determining the predicted class (as is visible in figure 4.10 (e)).

To simplify, in the next figures it will only be presented the original image, (a), the top-5 super-pixels, (c), and the heat-map, (e).

4.7.2.B Visualization and interpretation of the learned representations using LIME

Figure 4.11 shows the output obtained using the LIME algorithm for the different models. By analyzing this figure, it is possible to conclude that, for the same input image, both SSL pre-trained models look at different parts of each lesion (first and second rows of figure 4.11 a), b), c) and d)). This was already confirmed in Section 4.6.2.B. Additionally, it is also possible to verify that the model, resulted from early fusing the features of both SSL techniques, also looks at different aspects of the image and, combines the learned information from both models (last row of figure 4.11 a), b), c) and d)). **Looking at figure 4.11 it is visible that the fused model was more precise in highlighting the skin lesion since the weights given for the Rotation and the SimCLR when combined resulted in higher importance in the lesion part.**

As expected, this qualitative assessment proved that the fused model, apart from having higher performance, was also more accurate in detecting the different skin lesions. Therefore, this proved that the learned information of both SSL pre-trained models is, in fact, complementary. However, the question that arises is: 'Apart from being complementary is the combined information sufficient to avoid some of the limitations highlighted in section 4.6.2.B?

Figure 4.12 shows four examples of skin lesions where both the SimCLR or the Rotation pre-trained models had difficulties in detecting the skin lesion.

Looking at figure 4.12 (a) and (b) it is visible that the rotation technique had difficulties in detecting the skin lesion. However, the combination of both methods proved to overcome this limitation. In figure 4.12 (c) and (d) it is possible to confirm that since both lesions have less contrast than the margin and since they are quite symmetrical, both SSL pre-trained models had difficulties in detecting the lesion. However, when the features are combined the importance weights tended to highlight the lesions.

Therefore, by combining both models some of the limitations presented in both SSL pre-trained models could be avoided.

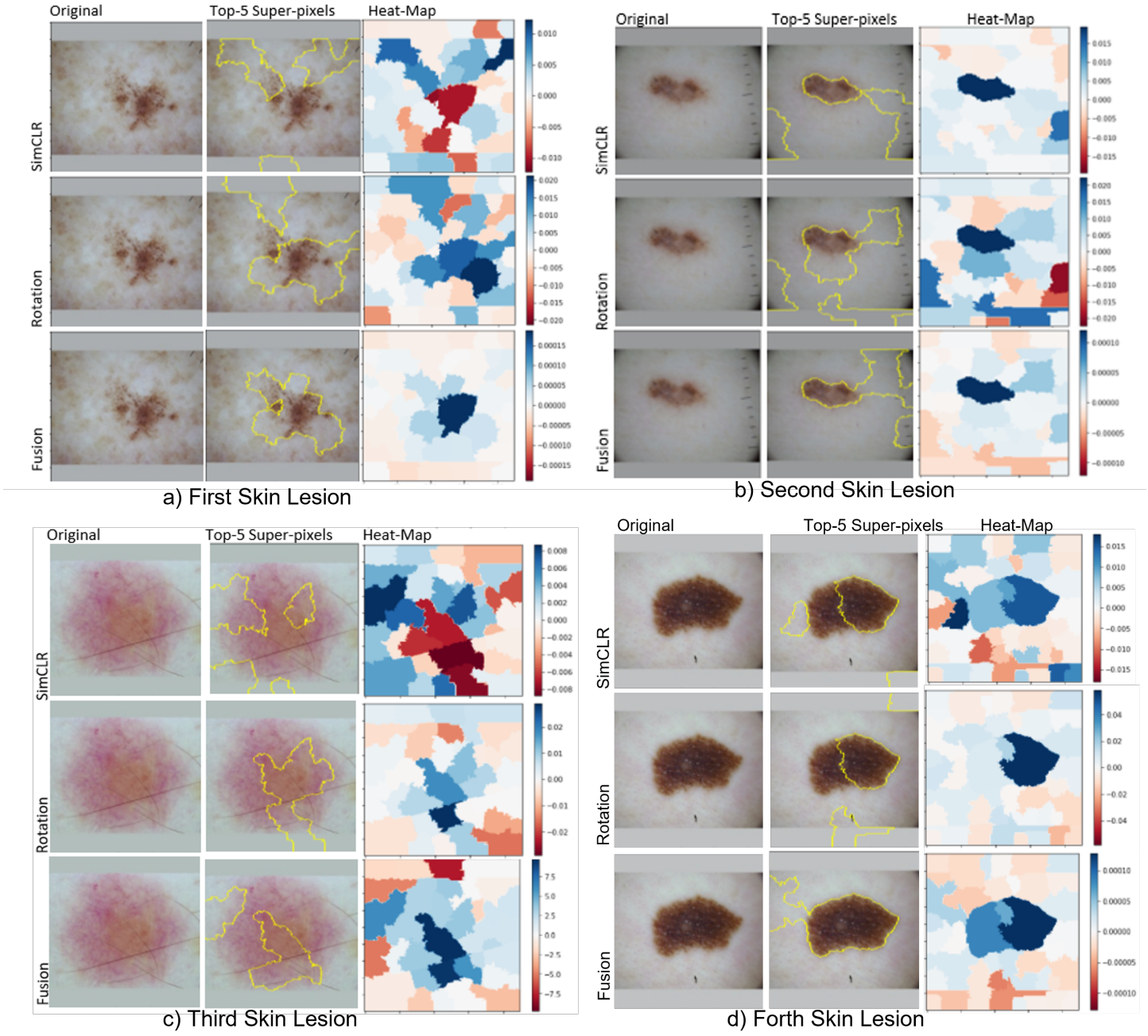
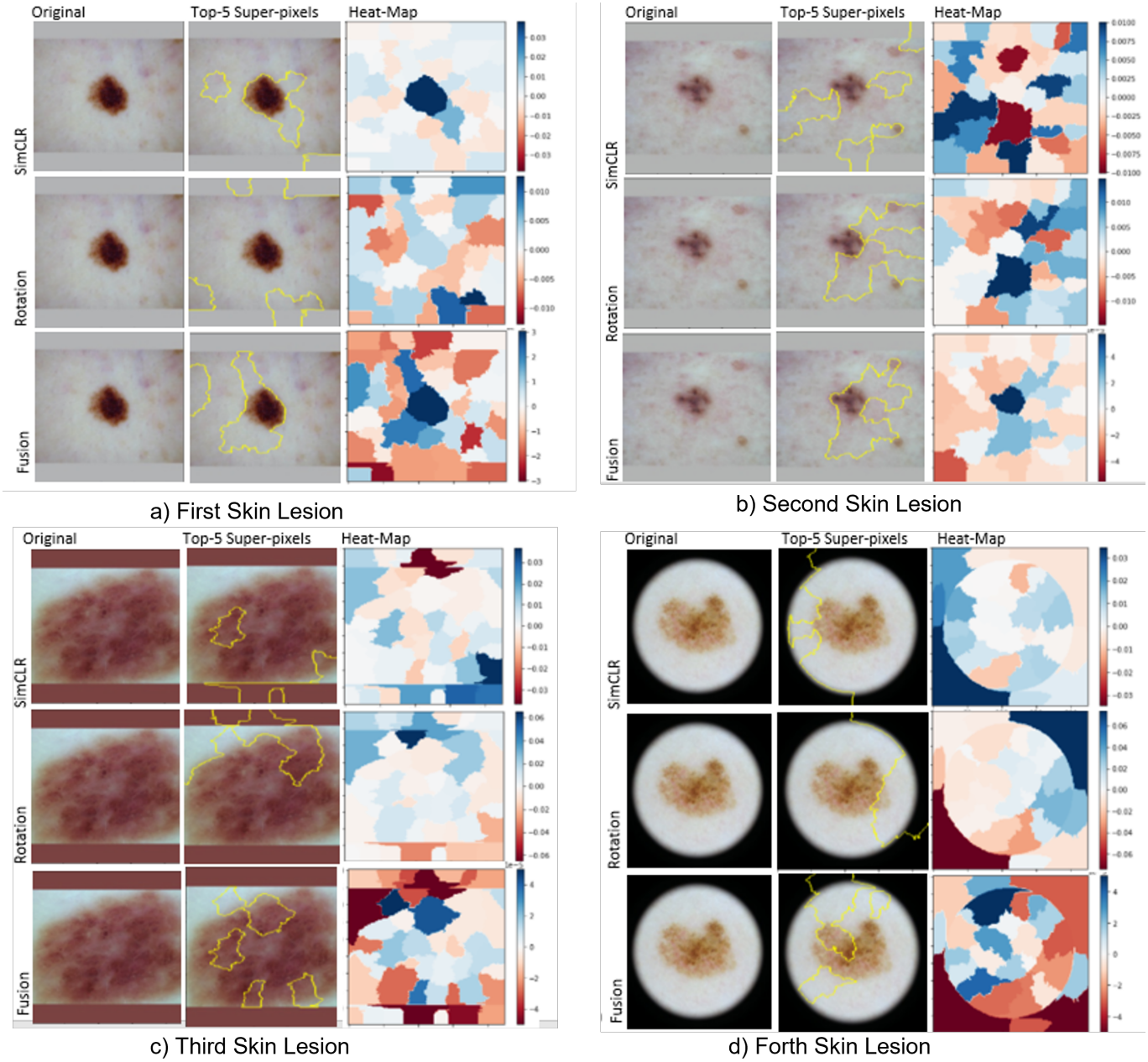


Figure 4.11: Example of different lesion visualizations using the LIME algorithm for each SSL pre-trained model (Rotation, SimCLR and Early Fusion) for partition 1.



a) First Skin Lesion

b) Second Skin Lesion

c) Third Skin Lesion

d) Forth Skin Lesion

Figure 4.12: Example of different lesion visualizations using the LIME algorithm for each model SSL pre-trained model (Rotation, SimCLR and Early Fusion) for partition 1. This figur highlights the limitations of both SSL methods.

4.8 Further Quantitative Evaluation of all initialization techniques

In order to better analyze the results obtained in table 4.5 and table 4.6 it was performed a statistical significance test which is presented in the appendix A.

To corroborate the conclusions made by analyzing table 4.5 and table 4.6 a boxplot was implemented.

Figure 4.13 presents the boxplot containing all different initialized models (minus the late fusion since it had a worse performance).

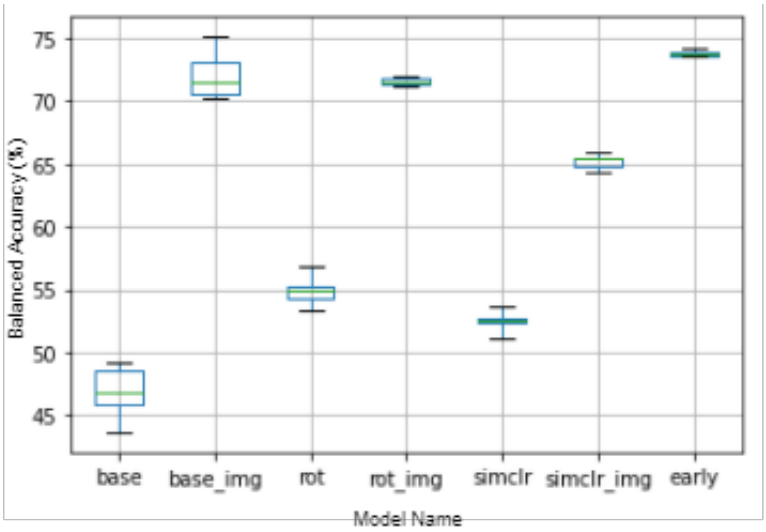


Figure 4.13: Boxplot of the different implemented models. The green line represents the median and the box represents the middle 50% of all data points, which represent the core region where the data is situated. The baseline models are written as 'base', the rotation as 'rot', the early fusion model as 'early' and the models fine-tuned with ImageNet weights end with 'img' in their name.

Looking at figure 4.13 it is possible to confirm that the model with the highest accuracy and stability is the early fusion. This model gathers both learned features from the Rotation and SimCLR pre-trained models and it confirms that this learned information is complementary. The Rotation (rot_img) model has similar accuracy as the baseline model pre-trained in ImageNet (base_img), however, it is more stable. Both SimCLR (simclr and simclr_img) and Rotation (rot and rot_img) models have higher stability than the baseline (base and base_img), this is possible to confirm since the box is narrower for both self-supervised models.

4.8.1 State-of-the-Art comparison

As mention in section 2.3.3, SSL has been used in the skin image context. Both Li *et al.* [24] and Tajbakhsh *et al.* [29] applied SSL techniques with color-based pretext tasks to the segmentation of skin lesions. Kwasigroch *et al.* [23] applied two SSL techniques based on geometric distortion to the skin cancer classification task. The closest work to the one executed in this thesis is that of Chaves *et al.* [50], in which they assess five SSL contrastive techniques against a competitive supervised baseline and conclude that SSL is competitive both in reducing variability and improving model accuracy. Despite the promising results, it is still unclear which is the best SSL strategy for skin images and all works focus solely on a quantitative analysis, disregarding the impact of SSL on the features learned by the model.

Therefore, to better compare this thesis trained models with the classification ones presented in table 2.2, the AUC score was implemented. The results obtained are shown in table 4.7.

Table 4.7: Evaluation of the different models using the AUC score.

Initialization	Technique	AUC (%)
Imagenet + SSL	Baseline	94,64 ± 0,31
	Rotation	94,72 ± 0,25
	SimCLR	92,92 ± 0,12
	Early Fusion	**94,94** ± 0,22

Analyzing table 4.7 it is possible to confirm, that both the Rotation and the Early fusion model have better results than the baseline. In addition, as seen before in section 4.6 it is also visible that all three models have lower variability than the baseline model.

To ease the state-of-the-art comparison, the results were reunited in table 4.8. This table presents this thesis AUC score as well as both the Kwasigroch *et al.* [23] and the Chaves *et al.* [50] results (recall table 2.2). It is important to recall that all three works have been trained using different datasets with different purposes. The ISIC 2017 [2] task had the objective of differentiating two classes - malignant (MEL) and benign (NV and BKL). The ISIC 2020 [68] had the same purpose, but it included more lesions within each class: benign (NV, atypical melanocytic proliferation, café-au-lait macule, lentigo NOS, lentigo simplex, solar lentigo, lichenoid keratosis and BKL) and malignant (MEL).

Table 4.8: Evaluation of the different models using the AUC score.

Authors	Dataset	Technique	AUC (%)
Kwasigroch et al., 2020 [23]	ISIC 2017	Jigsaw	83,4
		Rotation	84,2
Chaves et al., 2021 [50]	ISIC 2020	BYOL	94,6 ± 0,5
		InfoMin	94,4 ± 0,5
		MoCo	93,9 ± 0,7
		SimCLR	**95,6 ± 0,3**
		SwAV	95,3 ± 0,6
Thesis work	ISIC 2019	Baseline	94,6 ± 0,3
		Rotation	94,7 ± 0,2
		SimCLR	92,9 ± 0,1
		Early Fusion	**94,9 ± 0,2**

Looking at table 4.8, it is possible to confirm that the results presented in this thesis have higher AUC score than the ones presented in the Kwasigroch *et al.* [23] work. In addition, looking at the scores obtained in the Chaves *et al.* [50] work, it is visible that this thesis best work, which is the early fusion model, presented a better performance than most models (Sup. Baseline, BYOL, InfoMin, and MoCo). However, the early fusion model showed to have lower score than both the SimCLR (-0.66%) and the

SwAV (-0.36%) models. Analyzing the standard deviation it is possible to conclude that the results obtained in this thesis show even less variability than the ones presented in the Chaves *et al.* [50] work.

4.9 Complementary Study: Study the impact of adding more data to the SSL pre-trained models

SSL is known to benefit from using more data. In the pre-training phase this technique does not use labels, therefore the performance of the network increases with the variability of the available data. The more data, the more accurate the model can be to execute the intended SSL technique. It is also important to recall that depending on the level of difficulty of the task, the more it benefits from using more data. For example, when applying a contrastive task the network gathers the features which are similar and repels the different ones, meaning that this is quite difficult for a network to do. However, when using simpler tasks, such as geometric distortion, the network would benefit less from using more data that when applying contrastive learning tasks.

The impact of adding more data on the SSL pre-trained models was studied. It was opted to add 50% more data (using the ISIC 2020 dataset [68]). To compare this complementary study with the one executed in section 4.7.1, table 4.9 was created.

Table 4.9: Application of the Monte Carlo Sampling using more 50% of unlabeled data.

SSL Dataset	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
ISIC 2019	Rotation	71,47 ± 0,30	62,37 ± 0,74	65,7 ± 0,47	95,77 ± 0,05
	SimCLR	65,51 ± 0,55	54,47 ± 2,71	58,28 ± 1,95	95,17 ± 0,18
	Early Fusion	73,78 ± 0,24	68,41 ± 2,07	70,99 ± 2,61	96,40 ± 0,36
50% more data	Rotation	70,22 ± 0,98	62,89 ± 1,56	66,04 ± 0,96	95,73 ± 0,39
	SimCLR	**67,48** ± 0,58	**64,34** ± 6,05	**65,16** ± 3,69	**95,44** ± 0,63
	Early Fusion	**74,28** ± 0,58	**71,15** ± 1,57	**73,03** ± 0,96	**96,41** ± 0,15

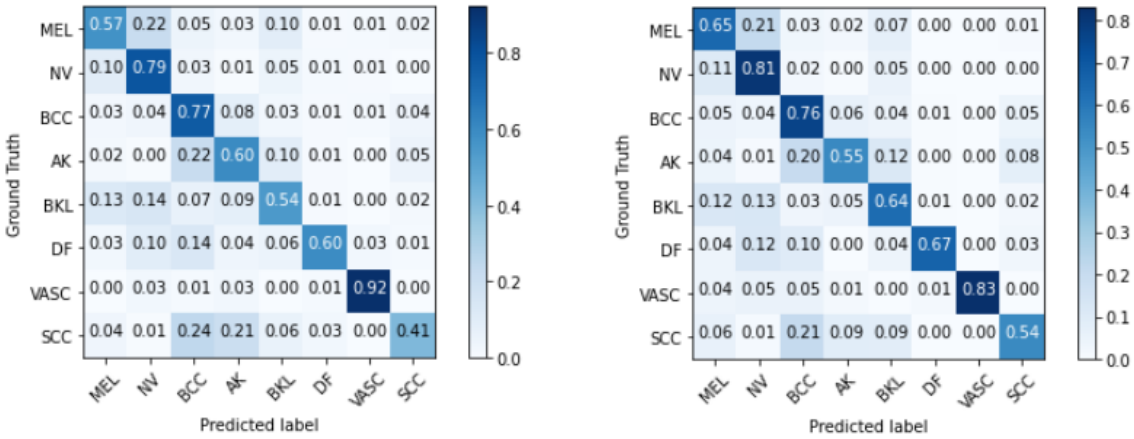
Analyzing table 4.9 it is possible to conclude that, in fact, the SimCLR task benefited from the use of more data. On the other hand, the Rotation technique had similar metrics to the previous training. This could be explained by the fact that this is a simpler task. The fusion of both techniques also showed to benefit with the use of more data, this was expected since the SimCLR also improved.

4.9.1 Differences in the predicted classes using the SimCLR technique

In order to understand the impact of adding more data while using the SimCLR technique, the different confusion matrices, obtained through the distinct models, were compared: using ISIC 2019 dataset, using 50% more data.

First, it is important to highlight that the common lesions between the ISIC 2019 and 2020 dataset [68] are only NV, BKL (benign) and MEL (malignant) lesions. Therefore, the results presented in this section could have been better if the ISIC 2020 dataset contained more lesions from the different classes presented in the ISIC 2019 dataset. It was opted to use the 2020 dataset since the ISIC 2019 used images from previous challenges.

Figure 4.14 shows two different confusion matrices regarding the validation set obtained for one of the partitions.



(a) Using the ISIC 2019 dataset

(b) Using 50% more data

Figure 4.14: Confusion matrices obtained for the SimCLR model for the best partition.

The diagonal of the matrix represents the sensitivity by class and ideally would be 1. This would mean that all classes were classified correctly. The remaining entries represent the misclassifications. Looking at figure 4.3 (a) and (b) it is possible to confirm that the classes which benefited most from the addition of 50% more data were: the SCC (+13%), BKL (+10%), MEL (+8%), DF (+7%) and finally NV (+2%).

Looking at all partitions it was possible to see similar results, however, the most common conclusion was the improvement of the melanoma classification, this could be explained by the fact that this 2020 dataset is highly imbalanced. This means that there was mainly an improvement in the benign classes (BKL, DF and NV) and the MEL class, which could be explained by the fact that the ISIC 2020 only contained benign and MEL lesions.

Appendix A contains the study executed with the use of 100% more data.

4.10 Final Evaluation in the Test Set

In order to verify how well the models obtained in this thesis generalized, it was opted to evaluate them using the test set provided by the ISIC 2019. This is an independent set without ground truth data, and the evaluation of the models was performed on an online platform [69].

To compare the results obtained using the test dataset [2], the ISIC leaderboard [12] was analyzed. The classification in this challenge is based on the weighted accuracy of all classes (weighted average of the SE). It is important to recall that the test dataset contains a class unknown. However, in this thesis, it was opted to use the BACC score (without taking into account the class unknown), since the same importance is given to all the classes, even if they contain a different number of examples.

Table 4.10 contains the performance achieved by the different initialization models implemented in the validation and held-out test set for the best partition. The accuracy containing the class unknown is presented in the column 'Test w/ UNK class' of table 4.10. However, the BACC without considering the unknown class is presented in the column 'Test' [3] of table 4.10.

Table 4.10: Evaluation of the different models using the test set.

SSL Dataset	Technique	BACC		
		Valid	Test w/ UNK	Test
ISIC 2019	Baseline	0,715	0,443	0,447
	Rotation	0,715	0,481	0,499
	SimCLR	0,656	0,444	*
	Early Fusion	0,738	0,446	0,483
50% more data	Rotation	0,712	0,451	0,478
	SimCLR	0,675	0,452	0,433
	Early Fusion	0,743	0,423	*

Analyzing table 4.10 it is possible to verify that both the SimCLR and the fused model increased their BACC score performance in the test set with the use of more data. This contributed to prove that the more data, the more accurate the model can be to execute the intended SSL technique depending on the level of difficulty of the task. Meaning that the Rotation technique showed little improvement since it is a simpler task than the SimCLR technique. It is also visible that the model with better results in the evaluation of the test set is the Rotation model. This could be explained by the fact that the pre-processing process of the dataset was made using padding of black margins, which as seen in section 4.6.2.B, can be a limitation of the SimCLR model. The model tended to focus more on the parts of higher contrast of the image, which in this case were the margins. Therefore, since the SimCLR had a worse performance the fusion of both models also had difficulties due to the higher contrast in the margins.

[2] The pre-processing of the test set instead of adding the most predominant color of the image (as shown in section 3.2.1), it was opted to add black margins, due to time restrictions.

[3] The entries containing '*' were not presented in the top 200 of the online platform and, therefore, the SE score was not available. Meaning that the BACC could not be calculated.

Additionally, it is interesting to verify, as seen before in section 4.6.1 and 4.7.1, that both the Rotation and the Early Fusion models had a higher performance than the baseline model.

5

Conclusions and Future Work

Contents

5.1 Conclusions

This thesis performed a robust assessment of the impact of SSL as a pre-training technique for skin cancer diagnosis. In particular, it performed a quantitative and qualitative analysis of the different pipelines. During this assessment, two SSL techniques were compared: Rotation and SimCLR. The experimental results show that there are benefits while using SSL. It was possible to observe that when applying these techniques, the classification CNN appeared to have more stability in its performance. It is beneficial to have models that are more stable since this means they are more trustworthy to apply to other data. Additionally, this proved that when combining transfer learning with SSL, the generalization problem that occurs when using TL is filtered. TL uses natural images that have a different domain to the skin lesion ones. Therefore the network resulted from applying transfer learning, will have neurons that remain loyal to the natural images. By applying SSL these neurons are 'corrected' and the obtained network generalizes better to the skin lesion images. This is believed to be the first work that provided a qualitative analysis of the features learned by the SSL strategies. This study led to the conclusion that each model learned different information from the data. Additionally, it was also possible to conclude that each SSL technique had some limitations: the Rotation had difficulties in detecting symmetrical lesions, while the SimCLR, as some images contained margins with higher contrast, tended sometimes to focus more on the margins than the lesion itself.

In order to verify if the information learned by both SSL models was complementary, it was studied the combination of both techniques that resulted in the highest performance (BACC $= 73,78 \pm 0,24\%$). In addition, it was also possible to conclude that the model resulted from combining both SSL techniques overcame some limitations that each SSL model had individually.

As SSL is known to benefit from using more unlabeled data, it was also studied the impact of adding 50% more data to the SSL pre-trained models. It was possible to observe that depending on the level of difficulty of the task, the more the model benefits from using more data. Therefore, the SimCLR task benefited more from the increase of data, since this is a more challenging task when compared to the Rotation. The fusion of both techniques also showed to benefit with the use of more data, this was expected since the SimCLR also improved.

Finally, the pre-trained models were evaluated using the test set. This study reinforced the conclusion that the SimCLR model trained using more data had higher capability to generalize to new data. Additionally, the Rotation and the Early fusion models have also shown to have higher performance than the baseline model even in the test set.

5.2 Future Work

The results obtained in this thesis highlighted the importance of using SSL techniques. However, there is room to improve the results. Therefore, some points can be highlighted regarding some topics that can be studied in future works:

- During this thesis for the SimCLR there were used three combinations of image transformations: horizontal flips, central crops and rotations (0, 90, 180, or 270 degrees). The impact of random color distribution and random gaussian blur were also evaluated, however, these experiments resulted in a lower performance of the model. In the future it could be interesting to try less 'aggressive' transformations such as normalizing each image color, which could result in better performance;
- Instead of combining both models using early fusion, it could be interesting to train a single network to execute both SSL techniques simultaneously and, therefore, its final performance could have better results;
- It could be interesting to also try another SSL technique and combine it with the SimCLR and the Rotation technique using the Early Fusion method. Therefore, since the Rotation technique focused more in the structure of the lesion and the SimCLR in the contrast, it could be beneficial to implement in the future a technique related to color such as the ColorMe [24] technique.

Bibliography

[1] 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*, vol. 5, 08 2018.

[2] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. 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)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 168–172.

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

[4] "University freiburg: Self-supervised learning, 2020." [Online]. Available: https://rl.uni-freiburg.de/teaching/ss20/selfsupervisedlearning

[5] "Convolutional neural network for cell classification using microscope images of intracellular actin networks, 2020." [Online]. Available: https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0213626.g002

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

[7] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734–1747, 2016.

[8] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1422–1430.

[9] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 649–666.

[10] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, 2016.

[11] "Google ai, 2020." [Online]. Available: https://ai.google/

[12] "Isic archive, 2020." [Online]. Available: https://www.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D%20Que%20tem%20ime

[13] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conference*, vol. 2004, no. 1. Society for Imaging Science and Technology, 2004, pp. 37–41.

[14] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," in *International Conference on Learning Representations (ICLR)*, Vancouver, Canada, Apr. 2018. [Online]. Available: https://hal-enpc.archives-ouvertes.fr/hal-01864755

[15] A. Chaudhary, "The illustrated simclr framework, 2020." [Online]. Available: https://amitness.com/2020/03/illustrated-simclr/

[16] "Understanding how lime explains predictions, 2018." [Online]. Available: https://towardsdatascience.com/understanding-how-lime-explains-predictions-d404e5d1829c

[17] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114 822–114 832, 2020.

[18] *Euromelanoma: Cancro da Pele, 2020*. [Online]. Available: https://www.euromelanoma.org/portugal/cancro-da-pele

[19] "Skincancer foundation, 2020." [Online]. Available: https://www.skincancer.org/early-detection/

[20] Y. Fujisawa, S. Inoue, and Y. Nakamura, "The possibility of deep learning-based, computer-aided skin tumor classifiers," *Frontiers in medicine*, vol. 6, p. 191, 2019.

[21] "Dermoscopy, 2020." [Online]. Available: https://www.the-dermatologist.com/content/dermoscopy

[22] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2051–2060.

[23] A. Kwasigroch, M. Grochowski, and A. Mikołajczyk, “Self-supervised learning to increase the performance of skin lesion classification,” *Electronics*, vol. 9, no. 11, p. 1930, 2020.

[24] Y. Li, J. Chen, and Y. Zheng, “A multi-task self-supervised learning framework for scopy images,” *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 2005–2009, 04 2020.

[25] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, “Self-supervised learning for medical image analysis using image context restoration,” *Medical image analysis*, vol. 58, p. 101539, 2019.

[26] A. Taleb, C. Lippert, T. Klein, and M. Nabi, “Multimodal self-supervised learning for medical image analysis,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 661–673.

[27] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[28] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, “Knowledge transfer for melanoma screening with deep learning,” *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 297–300, Apr 2017.

[29] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding, “Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 1251–1255.

[30] Y. Lecun, Personal communication in 2020 AAAI Conference on Artificial Inteligence.

[31] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[32] X. Yang, X. He, Y. Liang, Y. Yang, S. Zhang, and P. Xie, “Transfer learning or self-supervised learning? a tale of two pretraining paradigms,” *ArXiv*, vol. abs/2007.04234, 2020.

[33] I. Namatevs, “Deep convolutional neural networks: Structure, feature extraction and training,” *Information Technology and Management Science (Sciendo)*, vol. 20, no. 1, 12 2017.

[34] G. K. Abraham, V. Jayanthi, and P. Bhaskaran, “Convolutional neural network for biomedical applications,” in *Computational Intelligence and Its Applications in Healthcare*. Elsevier, 2020, pp. 145–156.

[35] S. Chan, V. Reddy, B. Myers, Q. Thibodeaux, N. Brownstone, and W. Liao, "Machine learning in dermatology: Current applications, opportunities, and limitations," *Dermatology and Therapy*, vol. 10, 04 2020.

[36] "Isic challenge 2019 leaderboards, 2020." [Online]. Available: https://challenge.isic-archive.com/leaderboards/2019

[37] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[38] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, 12 2017.

[39] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, *Chapter 1 - Analytics Defined*. Boston: Syngress, 2015, pp. 1 – 12.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 1097–1105.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[42] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[44] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. M. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting," *CoRR*, vol. abs/1808.01694, 2018.

[45] Y. Pan and Y. Xia, "Residual network based aggregation model for skin lesion classification," *CoRR*, vol. abs/1807.09150, 2018.

[46] K. M. Li and E. C. Li, "Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks," *arXiv , vol. abs/1807.08332*, 2018.

[47] R. M. Steven Zhou, Yixin Zhuang, "Multi-category skin lesion diagnosis using dermoscopy images and deep cnn ensembles," In arXiv. 2019.

[48] J. Zhuang, W. Li, S. Manivannan, R. Wang, J. J.-G. Zhang, J. Pan, G. Jiang, and Z. Yin, “Skin lesion analysis towards melanoma detection using deep neural network ensemble,” *ISIC Challenge 2018*, vol. 2, 2018.

[49] V. Chouhan, “Skin lesion analysis towards melanoma detection with deep convolutional neural network,” In arXiv. 2019.

[50] L. Chaves, A. Bissoto, E. Valle, and S. Avila, “An evaluation of self-supervised pre-training for skin-lesion analysis,” *CoRR*, vol. abs/2106.09229, 2021.

[51] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” *In European Conference on Computer Vision, Germany*, pp. 65–84, 2016.

[52] M. Noroozi, H. Pirsiavash, and P. Favaro, “Representation learning by learning to count,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5899–5907.

[53] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *CoRR*, vol. abs/1511.06434, 2016.

[54] J. Donahue, P. Krähenbühl, and T. Darrell, “Adversarial feature learning,” *CoRR*, vol. abs/1605.09782, 2016.

[55] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[56] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[57] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent: A new approach to self-supervised learning,” in *Neural Information Processing Systems, Montréal, Canada. hal-02869787v2*. Elsevier, 2020.

[58] M. Laskin, A. Srinivas, and P. Abbeel, “Curl: Contrastive unsupervised representations for reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 5639–5650.

[59] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct 2019.

[61] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you? explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[62] K. Liu, Y. Li, N. Xu, and P. Natarajan, “Learn to combine modalities in multimodal deep learning,” *CoRR*, vol. abs/1805.11730, 2018.

[63] I. Challenge, “Isic challenge datasets, 2020.” [Online]. Available: https://challenge.isic-archive.com/data

[64] T. Fawcett, “Introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, pp. 861–874, 06 2006.

[65] “Developer guides.” [Online]. Available: https://keras.io/guides/

[66] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *CoRR*, vol. abs/1603.04467, 2016.

[67] “Welcome to colaboratory.” [Online]. Available: https://colab.research.google.com/

[68] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman *et al.*, “A patient-centric dataset of images and metadata for identifying melanomas using clinical context,” *Scientific data*, vol. 8, no. 1, pp. 1–8, 2021.

[69] “Isic challenge.” [Online]. Available: https://challenge.isic-archive.com/

[70] L. Bai and D. Kalaj, “Approximation of kolmogorov–smirnov test statistic,” *Stochastics*, vol. 93, no. 7, pp. 993–1027, 2021.

Extra Information

A.1 SSL applied to the medical image analysis

As mentioned in section 2.3.1, in order to better understand what was being done with the SSL techniques and since there were few works in the skin cancer field, it was opted to analyze the implemented techniques in the context of medical image analysis. Table A.1 shows different works that applied SSL techniques to different medical applications. However, it is important to stress that most self-supervised techniques are very recent and, consequently, there are still few works that use them.

Table A.1: Application of self-supervised learning to medical diagnosis.

Authors	Goal	Features	SSL	TL	Scratch	TL + SSL
Chen et al., 2019 [25]	Fetus Classification	Context Restoration (CR)	87,56	-	-	-
	Abdominal Multi-organ Localization	Context Restoration (CR)	5,99 $\pm$ 9,83	-	-	-
	Brain Tumour Segmentation	Context Restoration (CR)	85,57	-	-	-
Tajbakhsh et al., 2019 [29]	Lung Lobe Segmentation	Rotation	0,94	-	0,92	-
	DR Classification	Rotation	0,75	0,71	0,70	-
	FPR Nodule Detection	3D Patch reconstruction	0,72	-	0,71	-
Li et al., 2020 [24]	Cervix type Classification	ColoMe	61,65	57,39	62,50	65,91

Analyzing table A.1, it was possible to see that the application of self-supervised learning to medical image analysis ([25] [29] [24]) can lead to better performance when comparing to training from scratch or transfer learning [29]. However, this may not always happen since there are a variety of self-supervised techniques and each technique can affect positively or negatively the performance depending on the goal task. For example, when segmenting the lung lobe [29] the rotation technique [14] was applied, once there is a consistency in the thorax geometry. This way, by implementing random rotation on the images and by forcing the network to predict which rotation was applied, the neural network learns visual features that characterize the structure of the thorax. Hence, when applying the weights from this pretext task the segmentation improves when compared to pre-training from the ImageNet.

A.2 Statistical Significance test

As mentioned in section 4.8, here is presented the executed statistical significance test. Since the data did not have a Gaussian distribution it was not possible to apply the t-test. Therefore, the Kolmogorov-Smirnov [70] test was applied to see if the difference between the means of two distributions is statistically significant or not. In this test, if the p-value is very small, this suggests that the difference between the two populations is significant. The 'ks_2samp' python function was used.

To better compare the different methods there were executed 9 comparisons. First, within each method (baseline, SimCLR and Rotation), secondly within the models trained from scratch and, thirdly, between the models fine-tuned using ImageNet weights. Table A.2 reunites the results obtain for this statistical test.

Table A.2: Results obtained through the statistical significance test.

Initialization	1st Model	2nd Model	P-value	Conclusion
Within Each Method	Base	Base_img	0,0079	Diferrent distributions (reject Ho)
	SimCLR	SimCLR_img	0,0079	Diferrent distributions (reject Ho)
	Rot	Rot_img	0,0079	Diferrent distributions (reject Ho)
Scratch	SimCLR	Rot	0,079	Same distributions (fail to reject Ho)
	SimCLR	Base	0,0079	Diferrent distributions (reject Ho)
	Rot	Base	0,0079	Diferrent distributions (reject Ho)
ImageNet	SimCLR_img	Rot_img	0,0079	Diferrent distributions (reject Ho)
	SimCLR_img	Base_img	0,0079	Diferrent distributions (reject Ho)
	Rot_img	Base_img	0,873	Same distributions (fail to reject Ho)

Analysing table A.2 it is possible to confirm, as seen before in section 4.8, that both models trained from scratch: the SimCLR and the Rotation (Rot) have similar median value and both models fine-tuned with ImageNet: the Baseline (Base_img) and the Rotation (Rot_img) also have similar median value. This means, that these models are comparable.

A.3 Complementary Study: Study the impact of adding 100% more data to the SSL pre-trained models

As seen in section 4.9, SSL is known to benefit from using more data, in this section it was studied the impact of adding 50% more data on the SSL pre-trained models. Here, it was opted to verify the impact of adding 100% more data (using the ISIC 2020 dataset [68]). Table A.3 gathers the evaluation metrics obtained from training the previous models using 100%[1] more data.

Table A.3: Application of the Monte Carlo Sampling using 100% more of unlabeled data.

SSL Dataset	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
100% more data	Rotation	69,42	61,62	64,72	95,42
	SimCLR	69,66	61,47	64,81	95,49
	Early Fusion	76,28	74,35	75,25	96,2

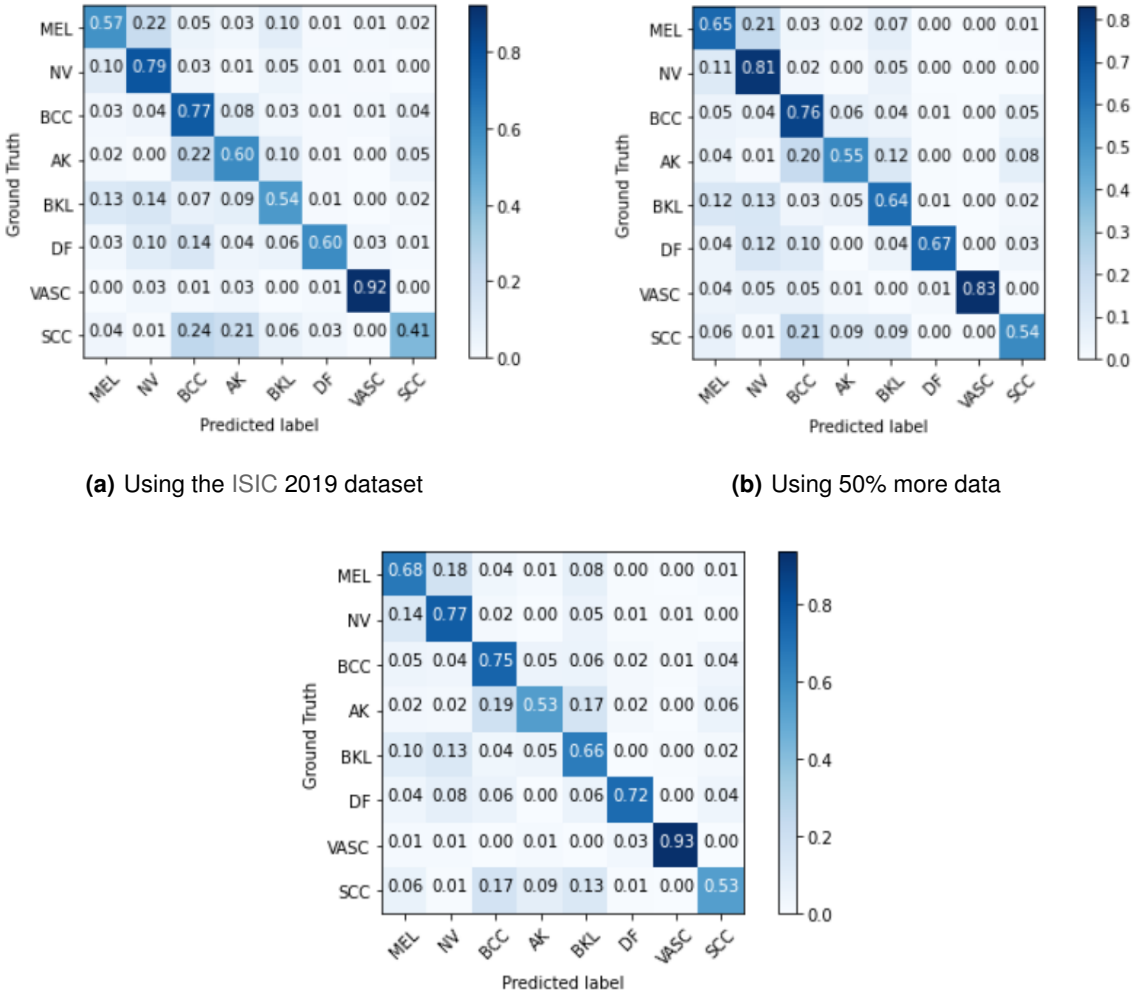
Despite having used only one fold, it is visible in table A.3 similar results as obtained in section 4.9. The SimCLR task benefited from the use of more data. The Rotation technique had similar metrics to the previous training and the fusion of both techniques also showed to benefit with the use of more data, this was expected since the SimCLR also improved.

[1] Due to time restrictions the results using 100% more data were obtained only from one fold.

A.3.1 Differences in the predicted classes using the SimCLR technique

In order to understand the impact of adding more data while using the SimCLR technique, the different confusion matrices, obtained through the distinct models, were compared: using ISIC 2019 dataset, using 50% and 100% more data.

Figure A.1 shows three different confusion matrices regarding the validation set obtained for one of the partitions.



(a) Using the ISIC 2019 dataset

(b) Using 50% more data

(c) Using 100% more data

Figure A.1: Confusion matrices obtained for the SimCLR model for the best partition.

As seen in section 4.9.1, looking at figure 4.3 (a) and (b) it is possible to confirm that the classes which benefited most from the addition of 50% more data were: the SCC (+13%), BKL (+10%), MEL (+8%), DF (+7%) and finally NV (+2%). However, analyzing figure 4.3 (a) and (c) it is visible that when adding 100% more data the classes which improved were: the SCC (+12%), BKL (+12%), MEL (+11%),

DF (+12%) and finally VASC (+1%).

B

ISBI 2022 Submission

ON THE IMPACT OF SELF-SUPERVISED LEARNING IN SKIN CANCER DIAGNOSIS

Maria Rita Verdelho and Catarina Barata

Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal

ABSTRACT

Deep neural networks (DNNs) are the standard approach for image classification. However, they require a large amount of data and corresponding annotations. Collecting medical data is a difficult task, due to privacy restrictions. Moreover, it is even harder to obtain the clinical labels, since these must be provided by specialists. Self-supervised learning (SSL) has emerged as a possibility to overcome this issue, since it uses non-annotated data to pre-train the DNN. Recently SSL has been applied in the context of skin cancer. However, the results were not conclusive. Moreover, a proper analysis of the impact of different SSL approaches is still missing. In this paper we investigate two SSL approaches: **Rotation** and **SimCLR**. Our results highlight the benefits of applying self-supervised learning to the classification of dermoscopy images. Additionally, we demonstrate that these approaches learn different and complementary features.

***Index Terms*—** Skin Cancer, Deep Learning, Self-Supervised Learning, Dermoscopy

1. INTRODUCTION

Skin cancer is one of the most common types of cancer worldwide [1]. In the past decade, the number of melanoma cases diagnosed has increased by 47% and in non-melanoma cancer about 5.400 people worldwide die every month due to this disease [2]. Skin cancer is also one of the most treatable forms of cancer when detected in an early stage. However, late detection can have a significant impact in mortality rates. Therefore, there is a need to develop a convenient and precise method to perform early diagnosis [3].

Over the past decade, deep neural networks (DNNs) have been developed to assist human experts and accelerate the process of skin cancer diagnosis [3]. However these methods require a huge amount of annotated data to obtain satisfactory results. Collecting medical data is a difficult task, due to privacy and law restriction, and it is even harder to obtain clinical annotations, since these must be provided by specialists [4]. To overcome this issue, the research community has been relying on transfer learning. This method consists of first training a model for a task using a large data base and then "recycle" it for a new target task [5]. These pre-trained models usually have deeper architectures than what is needed in medical image analysis [6]. Additionally, the color distribution of natural images is also very different from the medical ones [7], which can result in models that have difficulties in generalizing to the other data [6].

Self-supervised learning (SSL) has emerged as a strategy to avoid the annotation process. This technique takes advantage of unlabeled data to perform a pre-training of the DNN [8] [9], allowing the model to learn relevant image features that can later be applied to a specific task. Recently, SSL has been used in the skin image context. Both Li *et al.* [9] and Tajbakhsh *et al.* [6] applied SSL techniques with color-based pretext tasks to the segmentation of skin lesions. Kwasigroch *et al.* [4] applied two SSL techniques based on geometric distortion to the skin cancer classification task. The closest work to ours is that of Chaves *et al.* [10], in which they asses five SSL contrastive techniques against a competitive supervised baseline and conclude that SSL is competitive both in reducing variability and improving model accuracy. Despite the promising results, it is still unclear which is the best SSL strategy for skin images. Additionally, all works focus solely on a quantitative analysis, disregarding the impact of SSL on the features learned by the model.

This work aims to shed a new light on the application of SSL in the skin cancer context. Towards this goal we have develop a robust experimental framework to:

(i) investigate the impact of SSL on the training and generalization of a DNN for skin lesion diagnosis into 8 different classes, and demonstrate that even with a small dataset there are benefits in using SSL.

(ii) compare two different SSL approaches, one based on geometric distortion and another on contrastive learning.

(iii) for the first time provide a qualitative assessment of the impact of the different pre-training strategies, using explainability approaches.

(iv) demonstrate the complementarity of the features learned by the SSL strategies and the benefits of combining them.

To the best of our knowledge, this is the first work to perform a robust quantitative and qualitative validation of the impact of SSL, and to demonstrate the importance of combining different SSL techniques.

The remaining of the paper is organized as follows. Section 2 introduces the used methodologies, and Section 3 describes the experimental setup. Section 4 presents the results and Section 5 concludes the paper.

2. METHODOLOGIES

This section gives a brief overview of SSL and the two strategies adopted in this work, as well as the experimental setup adopted in the skin cancer problem.

2.1. Self-Supervised Learning (SSL)

SSL is a technique used to extract visual features from unlabeled data [4]. The main goal is to use the learned weights to initialize a DNN for a specific target task, which is, in the skin cancer image analysis, the classification of the different skin lesions. To achieve this goal, the model is trained to execute a pretext task, for which labels can be easily generated without human supervision.

Pretext tasks aims to extract different feature representations from the images. Therefore, it is important to select a SSL technique that is adequate to the wanted target supervised task. In this paper we will use two SSL techniques, which we believe to have a good performance on the skin image classification problem: Rotation [11] and the SimCLR [12].

2.1.1. Rotation technique

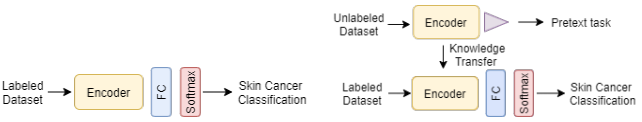
Rotation [11] is a classification-based technique, where the network is trained to predict which rotation (0º, 90º, 180º or 270º) has been applied to the image. Therefore, by predicting which rotation was applied to the input, the model is capable of extracting useful information from each image.

The training pipeline starts with a small set of geometric transformations, which will be applied to the dataset. Secondly, the transformed images are fed to the model and the DNN is trained to identify which rotation was applied to the original image. As mentioned before, the set of geometric transformations defines the classification task, meaning that if there are four rotations then it is a 4-class classification problem.

2.1.2. SimCLR technique

SimCLR [12] is a SSL approach that applies the concept of contrastive learning to infer feature representations from the unlabeled dataset. Feature representations are learned by maximizing the agreement between differently augmented views of the same image via a contrastive loss, which will also accentuate the dissimilarity among different images. The key idea is when comparing the multiple images using the contrastive objective, the representations of corresponding views are 'attracted' to one another and the others are 'repelled'.

SimCLR can be divided into four main steps: 1) Random transformations are applied to the input, in order to obtain a pair of two augmented images. 2) Each augmented image within the pair is sent to an encoder. 3) The output representations of the encoder are then sent to a multi-layer perceptron (MLP). 4) The contrastive loss is applied in the feature space given by the MLP.



(a) Supervised Learning. (b) Self-supervised Learning.

Fig. 1. Proposed framework using different initialization techniques applied to the skin cancer diagnoses. In both models the last layer is fully-connected one with 8 units. The triangle represents the last layers of the DNN specific of the pretext-task.

2.2. Experimental Framework

This paper aims to perform a robust assessment of the impact of SSL as a pre-training technique, to initialize the weights of a DNN for skin cancer diagnosis. To better understand the impact of SSL, we perform a systematic assessment, adopting the following pipeline:

(i) **Baselines** - two standard supervised learning strategies, where the weights of the DNN are initialized either at random (trained from scratch) or using a pre-trained model on ImageNet (fine-tuning).

(ii) **Scratch + SSL** - standard SSL methodology, where the weights of the DNN are initialized at random and refined using either the Rotation or the SimCLR technique.

(iii) **ImageNet + SSL** - a variant of the SSL approach, that aims to leverage the information from a model pre-trained on the ImageNet dataset. Here, we initialize the weights of the model used in the SSL phase from ImageNet and refine them using either the Rotation or SimCLR approach.

(iv) **Fusion** - fusion of the DNNs pre-trained using the Rotation and SimCLR techniques both at the feature (early fusion) and classification (late fusion) level.

Fig. 1 (a) describes the proposed generic approach for the application of supervised learning (baselines) and Fig. 1 (b) describes the proposed approach for the application of self-supervised learning. For the latter, the first step consists of pre-training the DNN using the chosen pretext task and, secondly, fine-tuning the parameters of model to the classification task (this time using labels), by recycling the encoder and adding a fully connected layer to output the 8 classes presented in our skin cancer dataset. In all our experiments, the encoder is a ResNet-50 [13].

3. EXPERIMENTAL SETUP

3.1. Dataset and Evaluation Metrics

All experiments were performed using the ISIC 2019 [14] [15] [16]. This dataset comprises 25,331 dermoscopy images,

divided into 8 lesions classes: Actinic keratosis (AKIEC), Basal cell carcinoma (BCC), Benign keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Nevus (NV), Squamous cell carcinoma (SCC) and Vascular (VASC). These labels are only used to train the classification models (recall Fig. 1). The images were collected at different medical centers (each center generated images with different sizes, color and aspect ratio). Therefore, it was necessary to pre-process them. This process compensated the color and allowed all the images to have the same size, while maintaining their aspect ratio. After having resized all the images to the desired size (224x224), we applied the color constancy algorithm Shades of Gray as it is proposed in [17].

In order to compare the different initialization approaches and assess their robustness, we adopted a 5-time Monte Carlo sampling strategy, where the ISIC 2019 dataset was partitioned five times into training (70%) and validation (30%) sets. Based on this, we report the median and standard deviation of the following metrics: Balanced Accuracy (BACC), Precision, F1-Score, and Specificity.

3.2. Network Training and Computational Environment

The experimental framework was implemented using Tensorflow/Keras and one NVIDIA Tesla K80 GPU [1]. All models were trained for 60 epochs, using early stopping and the Adam optimizer [18]. The batch size was set to 32. For SSL, the losses are the categorial cross-entropy for the rotation task and for the SimCLR it was used the NT-Xent loss (with $\tau = 0.1$). For this task, we transformed the input image using horizontal flips, central crops and rotations of 0, 90, 180 or 270 degrees. We also studied the impacts of random color distribution and random Gaussian blur, however these experiments resulted in a lower performance of the model. Both tasks had a initial learning rate of $\eta = 10^{-4}$, however the rotation had a reduction factor of 0.75 and the SimCLR a exponential decay of 0.96. To train the classifier, we adopted the weighted categorical cross-entropy loss, where the weights are set to the relative frequency of each class, in order to account for the unbalance. Here the learning rate was set to $\eta = 10^{-5}$, with a reduction factor of 0.75. In order to prevent over-fitting, we also used online data-augmentation (random flips and rotations of multiples of 90 degrees).

4. RESULTS

The results section is divided into three parts: i) a quantitative analysis, where we compare the different approaches taking into consideration the selected evaluation metrics; ii) a qualitative analysis that used the Grad-CAM technique [19] to convey a more interpretable analysis of the impact of the various initialization strategies in the features learned by the model; and iii) a quantitative analysis of the fusion of SSL strategies.

4.1. Quantitative Analysis

Table 1 summarizes the median and standard deviation of the scores obtained for the different initialization techniques. By looking at Table 1 it is possible to see that there are some benefits in using SSL when compared to the baseline supervised training. By looking at the baseline trained from scratch (row 1) and to both rows trained from scratch with self supervised learning techniques (row 3 and 4) it is visible that both SSL techniques presented higher median and lower standard deviations. This proves that when comparing models trained from scratch there is a tendency to have **higher accuracy** and **more stability** (the standard deviation has a lower value) in the **models** that use **SSL**. By looking at the models trained using the ImageNet weights - the baseline (row 2) and to both models that used the SSL techniques (row 5 and 6)- it is visible that the latter two have a higher stability (lower standard deviation) even though both had smaller or similar accuracy to the baseline. This proves that when comparing models trained with the ImageNet weights there is a tendency to **have more stability** in the models that use SSL. Finally, looking at the SSL pre-trained models (row 2, 3, 5 and 6) it is possible to see that the **rotation technique has a higher accuracy** when compared to the model initialized with the SimCLR technique.

A shared conclusion between our work, [4], and [10] is that, when using SSL pre-trained models, there is an outperformance in general terms, especially in variability.

4.2. Qualitative Analysis

We opted to execute a qualitative analysis, since we wanted to understand what each model saw differently and what it learned in order to make the diagnostic decisions. Therefore, to analyze the differences between the learned representations for each initialization technique the Grad-CAM [19] was used. This is a technique used for visualizing the features learned by the DNN and the regions of an image that activate a certain label. Figure 2 shows the Grad-CAM results for the different initialization techniques (fine-tuned with ImageNet weights).

Figure 2 proves that for the same input image all three models look at different parts of each lesion. Therefore, apart from having different performances each model seems to learn different information about each class of lesion. The SimCLR pre-trained model tended to focus more in the parts of the lesion that presented higher contrast, while the Rotation looked more at the structure of each lesion. The ImageNet pre-trained model, was the least intuitive to interpret, since its focus varied between lesion and skin. After, analyzing a set of different images it was possible to confirm that each

[1]The source code will be released upon acceptance of the paper

Table 1. Application of the Monte Carlo Sampling with different initialization techniques: training the model from scratch or fine-tuning with ImageNet weights; application of two self-supervised learning (SSL) techniques -Rotation and SimCLR - and fusion of both techniques.

Initialization	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
Baseline	Scratch	46,82 ± 2,00	35,37 ± 3,84	37,24 ± 4,64	92,89 ± 0,55
	ImageNet	71,48 ± 1,82	65,14 ± 2,78	67,93 ± 1,75	96,04 ± 0,12
Scratch + SSL	Rotation	54,92 ± 1,15	40,54 ± 1,84	43,19 ± 2,04	93,39 ± 0,18
	SimCLR	52,54 ± 0,86	44,62 ± 1,39	47,53 ± 0,96	93,94 ± 0,18
ImageNet + SSL	Rotation	71,47 ± 0,30	62,37 ± 0,74	65,70 ± 0,47	95,77 ± 0,05
	SimCLR	65,37 ± 0,55	54,47 ± 2,71	58,28 ± 1,95	95,17 ± 0,18
Fusion	Early Fusion	**73,78 ± 0,24**	**68,41 ± 2,07**	**70,99** ± 2,61	**96,40** ± 0,36
	Late Fusion (mean)	57,09 ± 2,19	50,28 ± 1,41	52,02 ± 1,08	94,24 ± 0,19

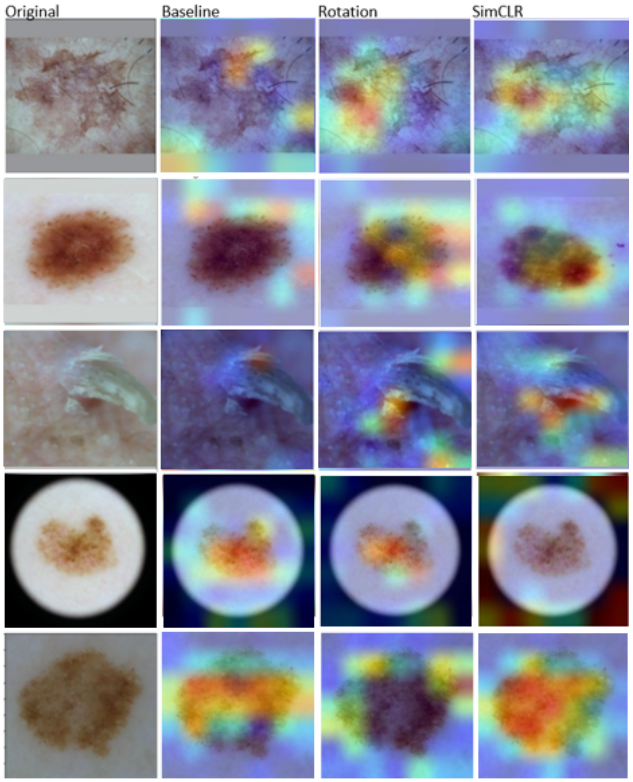


Fig. 2. Example of different lesion visualizations using the Grad-CAM algorithm (Baseline, Rotation and SimCLR).

method also had some limitations. The rotation had difficulties in detecting centered and symmetrical lesions, since each rotation of 90 degrees is similar, then the model does not learn useful information about this lesion. This limitation is visible in the fifth row and second column of fig. 2. The SimCLR showed to be more precise in detecting the lesion. However as some images contained margins with high contrast (black borders), this method tended to focus more on the margins than the lesion (exemplified in the fourth row and third column of fig. 2). Based on the qualitative results, the question that arose next was: **Is the information learned by both SSL techniques complementary?**

4.3. Fusion of SSL Approaches

As a consequence of the previous interrogation, we performed two tests that fused the models pre-trained with SSL. First, we used early fusion, which fuses the different methods in the feature space. Secondly, we used late fusion that fuses the models in the classification scores level (we applied the mean strategy). The fusion results appear in the last two rows of table 1. It is possible to conclude that the **early fusion (row 7) had better results than any other model both in stability and accuracy**, proving that in fact the features of both models have complementary information. However, the late fusion (row 8) proved to have worse results, meaning that the features are complementary, but not learned classification models.

5. CONCLUSIONS

This paper performed a robust assessment of the impact of SSL as a pre-training technique for skin cancer diagnosis. In particular, we performed a quantitative and qualitative analysis of the different pipelines. During this assessment we compared two SSL techniques: Rotation and SimCLR. Our experimental results show that there are benefits while using SSL. We observed that when applying this technique, the classification DNN appeared to have less variability in its performance. To the best of our knowledge this is the first work that provides a qualitative analysis of the features learned by the SSL strategies. This study led us to conclude that each model learned different information from the data. Therefore, we also studied the combination of the two SSL techniques which resulted in the highest performance. SSL is known to benefit from using more unlabeled data. Therefore, we plan to repeat both experiments using more unlabeled data in future work

6. REFERENCES

[1] *Euromelanoma: Cancro da Pele, 2020*.

[2] "Skincancer foundation, 2020," .

[3] Yasuhiro Fujisawa, Sae Inoue, and Yoshiyuki Nakamura, "The possibility of deep learning-based, computer-aided skin tumor classifiers," *Frontiers in Medicine*, vol. 6, pp. 191, 2019.

[4] Arkadiusz Kwasigroch, Michał Grochowski, and Agnieszka Mikołajczyk, "Self-supervised learning to increase the performance of skin lesion classification," vol. 9, pp. 1930, Nov 2020.

[5] Afonso Menegola, Michel Fornaciali, Ramon Pires, Flavia Vasques Bittencourt, Sandra Avila, and Eduardo Valle, "Knowledge transfer for melanoma screening with deep learning," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Apr 2017.

[6] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, Y. Xiao, Y. Lu, J. Liang, D. Terzopoulos, and X. Ding, "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 1251–1255.

[7] "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis*, vol. 58, pp. 101539, 2019.

[8] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2070–2079.

[9] Yuexiang Li, Jiawei Chen, and Yefeng Zheng, "A multi-task self-supervised learning framework for scopy images," *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 2005–2009, 04 2020.

[10] Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila, "An evaluation of self-supervised pre-training for skin-lesion analysis," *CoRR*, vol. abs/2106.09229, 2021.

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," In arXiv:1803.07728, 2018.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," In arXiv:2002.05709, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *In CVPR, 2015*.

[14] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, 08 2018.

[15] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan 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)," In arXiv:1710.05006, 2018.

[16] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy, "Bcn20000: Dermoscopic lesions in the wild," In arXiv:1908.02288, 2019.

[17] Graham Finlayson and Elisabetta Trezzi, "Shades of gray and colour constancy," 01 2004, pp. 37–41.

[18] Mohammed Alom, "Adam optimization algorithm," 06 2021.

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct 2019.