Deep learning applied to the classification of skin lesions

Giuliana Martins Silva

Dissertation presented to the School of Technology and Management of Bragança to obtain the Master Degree in Informatics. Work developed during the double degree exchange program between the Instituto Politécnico de Bragança (IPB) and the Universidade Tecnológica Federal do Paraná (UTFPR).

Work oriented by:
Professor PhD Fernando C. Monteiro
Assistant Professor PhD André E. Lazzaretti

Bragança
2023

Deep learning applied to the classification of skin lesions

Giuliana Martins Silva

Dissertation presented to the School of Technology and Management of Bragança to obtain the Master Degree in Informatics. Work developed during the double degree exchange program between the Instituto Politécnico de Bragança (IPB) and the Universidade Tecnológica Federal do Paraná (UTFPR).

Work oriented by:
Professor PhD Fernando C. Monteiro
Assistant Professor PhD André E. Lazzaretti

Bragança
2023

Dedication

To my family for all the support and motivation.

To my friends, for making these years more joyful and lighter.

To Enrico, my life partner, for everything we have been through and everything that is yet to come.

Acknowledgement

To Instituto Politécnico de Bragança (IPB) and Universidade Tecnológica Federal do Paraná (UTFPR) for the opportunity of being part of the exchange program between them.

To Prof. Dr. Fernando C. Monteiro and Prof. Dr. André E. Lazzaretti for all the help, support, and guidance provided throughout this work.

To the professors of IPB and UTFPR, and to each employee of IPB and UTFPR that were part of my academic development.

Abstract

Skin cancer has been a global health issue and its diagnosis is a challenge in the medical field. Among all the types of skin cancer, melanoma is the worst and can be lethal if not early treated. The use of deep learning techniques, specifically, convolutional neural networks can help to improve the accuracy and speed up the classification of skin lesions.

In this work, we aim to employ different image preprocessing techniques, various convolutional neural network models, data augmentation, and ensemble techniques to compare their results and provide an analysis of the data obtained. To achieve that, it was performed several experiments combining different image preprocessing techniques, which, paired with data augmentation strategies, aim to enhance the accuracy and reliability of the classification models. Additionally, three ensemble methods were tested to improve the classification systems' robustness and reliability by gathering the strengths of each model.

Our best result was the ensemble of EfficientNet-B2, EfficientNet-B5, and ResNeSt101 models with the application of data augmentation, and the combination of color constancy and hair removal techniques. This combined approach achieved a balanced accuracy of 0.8132.

By offering insights into the challenges faced, methodologies employed, and results obtained, this story aims to serve as a guide for researchers and practitioners aiming to advance the field of skin lesion classification using deep learning.

Keywords: Deep Learning; Skin Lesion Classification; Image preprocessing.

Resumo

O câncer de pele é um problema de saúde global e seu diagnóstico é um desafio na área médica. Entre todos os tipos de câncer de pele, o melanoma é o pior e pode ser letal se não tratado precocemente. O uso de técnicas de deep learning, especificamente, redes neurais convolucionais, pode ajudar a melhorar a precisão e acelerar a classificação de lesões de pele.

Neste trabalho, buscamos empregar diferentes técnicas de pré-processamento de imagens, vários modelos de redes neurais convolucionais, data augmentation e técnicas de ensemble para comparar seus resultados e fornecer uma análise dos dados obtidos. Para isso, foram realizados vários experimentos combinando diferentes técnicas de pré-processamento de imagens, que, combinadas com estratégias de data augmentation, visam melhorar a precisão e confiabilidade dos modelos de classificação. Além disso, três métodos de ensemble foram testados para melhorar a robustez e confiabilidade dos sistemas de classificação, reunindo os pontos fortes de cada modelo.

Nosso melhor resultado foi o ensemble dos modelos EfficientNet-B2, EfficientNet-B5 e ResNeSt101 com a aplicação de data augmentation e a combinação de técnicas de color constancy e remoção de pelos. Esta abordagem alcançou uma balanced accuracy de 0,8132.

Ao oferecer insights sobre as metodologias empregadas e resultados obtidos, este estudo visa servir como um guia para pesquisadores e profissionais que buscam avançar no campo da classificação de lesões cutâneas usando aprendizado profundo.
Palavras-chave: Aprendizado profundo; Classificação de lesões de pele; Pré processamento de imagem.

Contents

List of Tables

List of Figures

Acronyms

ACC accuracy.

AUC Area Under the ROC Curve.

CLAHE Contrast Limited Adaptive Histogram Equalization.

CNN Convolutional Neural Network.

DenseNet Dense Convolutional Network.

DL Deep Learning.

DNN Deep Neural Network.

DRN Deep Residual Network.

FC Fully connected Layers.

FPR False Positive Rate.

GAN Generative Adversarial Network.

GradCAM Gradient-weighted Class Activation Mapping.

HE Histogram Equalization.

IPB Instituto Politécnico de Bragança.

ISIC International Skin Imaging Collaboration.

PNASNet Progressive Neural Architecture Search.

PP preprocessing.

ReLU Rectified Linear Unit.

RNN Recurrent Neural Network.

SGD Stochastic Gradient Descent.

SVM Support Vector Machine.

TIMM PyTorch Image Models.

TNR True Negative Rate.

TPR True Positive Rate.

UTFPR Universidade Tecnológica Federal do Paraná.

VGG Visual Geometry Group.

Chapter 1

Introduction

The present thesis was developed with the aim of completing a Master's degree in Informatics. For this purpose, the topic proposed was the *Deep learning applied to the classification of skin lesions*. In this chapter, an introduction to the motivation behind this subject is provided. At the end of the chapter, the objectives and the structure of the document are presented.

1.1 Motivation

Nowadays, skin cancer is a global health problem. In 2020, the estimated number of new cases of skin cancer (melanoma and non-melanoma) worldwide was 1,522,708, corresponding to 7.89% of the total estimated cancer cases [1], [2]. In 2023, it is estimated that there will be 97,610 new cases of melanoma, the worst type of skin cancer in the US, occupying fifth place among the most incident types of cancer in the country [3].

This form of cancer arises predominantly due to prolonged, unprotected exposure to ultraviolet radiation from the sun, which can inflict serious damage to skin cells, including their DNA. The ensuing DNA damage may prompt genetic mutations, possibly instigating skin cancer progression. As the most prevalent cancer type, the incidence of skin cancer could be significantly reduced through preventive measures such as proper skin protection

against harmful solar rays [4].

Three primary types of skin cancer are recognized: basal cell carcinoma, squamous cell carcinoma, and melanoma. Melanoma is the most dangerous due to its potential to metastasize to other body parts, often leading to lethal outcomes if not detected and treated early. Therefore, precision in diagnosis and effective treatment are crucial for enhancing patient survival rates and quality of life [5].

Various methodologies exist for the preliminary evaluation of skin lesions, each designed to help identify malignant growths like melanoma at an early stage. Among these, the most popular is the ABCD criteria, which assess asymmetry, border irregularity, color variation and diameter greater than 6 mm. Another widely used method, the seven-point checklist, examines three major and four minor features, with a score of three or higher signaling the need for further scrutiny. Other techniques like the Ugly Duckling Sign, the Menzies method, and the Chaos and Clues approach focus on differentiating lesions based on comparative analysis and specific visual clues.

Associated with these methods, dermoscopy, an epiluminescence microscopy technique, enhances this diagnostic process by revealing skin structures not visible to the naked eye. It involves the use of a handheld device equipped with a magnifying lens and light source to closely examine lesions after applying a liquid medium like alcohol or oil to reduce surface reflection.

Compared to a naked-eye examination, dermoscopy can significantly improve diagnostic accuracy for skin cancers, especially melanoma. Despite the advantages offered by this approach, its efficacy is largely contingent on the expertise of the medical professionals utilizing it. Without proper training, there could be inconsistencies in how the diagnosis criteria are defined and applied [6].

With all the challenges involved in providing a precise and fast diagnosis, in the last decades, an effort has been made to create new technologies and tools to help with disease detection and treatment. The use of artificial intelligence, big data, and other technologies are being explored to develop sophisticated and accurate systems for disease diagnosis [7].

Skin lesion classification can greatly benefit from the use of artificial intelligence and

deep learning. Utilizing these technologies can provide a more accurate and effective way of identifying patterns and characteristics of skin lesions, extracting features, and classifying data. By learning from data, deep learning models can offer a promising opportunity for improvement in this field.

Leveraging the capabilities of deep learning, Convolutional Neural Network (CNN) has become a cornerstone in the fields of computer vision and medical image analysis. These specialized neural networks excel at recognizing and classifying different elements within images. Comprising multiple layers of interconnected nodes or neurons, CNNs automatically learn to identify key features and patterns in the data. This ability makes them ideal for tasks that involve high-dimensional data and intricate patterns, such as skin lesion classification. They have the capacity to automatically and adaptively learn from data, fine-tuning their internal parameters to make accurate predictions or classifications.

1.2 Objectives

In this work, we aim to investigate the use of deep learning techniques, image preprocessing, and data augmentation to classify skin lesions into three types: Melanoma, Seborrheic keratosis, and Nevi.

In our research, we explore the relationship between different methods and the final classification of skin lesions. We chose to work with 13 models, three ensemble approaches, and four image preprocessing techniques. We aim to provide a comprehensive evaluation of how these factors interact and influence the accuracy of the classification process.

Our experimentation will make use of the International Skin Imaging Collaboration (ISIC) 2017 Challenge dataset [8], which serves as a gold standard in the domain of dermatological image analysis. In addition to proposing a system for skin lesion classification, our research aims to answer some key questions: Does the incorporation of image preprocessing techniques and data augmentation methods significantly enhance classification accuracy? Do multiple CNN models, when used in conjunction, outperform single-model

setups? Lastly, which ensemble technique is most effective in aggregating the strengths of individual CNN models for optimal results?

Through this analysis, we seek to advance the discussion in the field of skin lesion classification and to provide insights about the enhancement in the accuracy and reliability of dermatological diagnoses using deep learning techniques. The source code of this project is available on GitHub [1].

1.3 Publications

This work has generated a paper entitled "Deep Learning Techniques Applied to Skin Lesion Classification: A Review" that was presented in the *2022 International Conference on Machine Learning, Control, and Robotics (MLCR)* [9], held in Suzhou, China, 29-31 October 2022, and published in the proceedings of the conference. Another paper entitled "An Evaluation of Image Preprocessing in Skin Lesions Detection" was presented in the *OL2A: International Conference on Optimization, Learning Algorithms and Applications 2023* [10], held in Ponta Delgada, Portugal, 27–29 September 2023, is not published yet. The last paper, entitled "Deep learning techniques applied to skin lesion classification" submitted and accepted in the *The 2023 2nd International Conference on Machine Learning, Control, and Robotics (MLCR 2023)* [11], held in Nanjing, China, 9-11 December 2023.

1.4 Document Structure

The document is structured as follows: The foundational concepts related to skin lesions, traditional diagnostic methods, and an explanation of deep learning concepts are presented in Chapter 2. Chapter 3 provides an overview of existing research, summarizing past methodologies. Chapter 4 details the research's specific methodologies and

[1] https://github.com/giuzis/masters

strategies, including an explanation of the dataset, image preprocessing techniques, data augmentation methods, ensemble techniques, and models used. The subsequent Chapter 5 showcases the outcomes of the research, providing a thorough analysis of results and comparisons between methods. Finally, the document concludes with Chapter 6, which encapsulates the main findings, discusses their implications, and potentially points toward future research directions in skin lesion classification using deep learning.

Chapter 2

Background

In this chapter, we set the context required to understand the problem and approaches used in this work. This chapter is divided into three key sections: Skin Lesions, Deep Learning, and Skin Lesion Datasets. In the first section, we delve into the three types of skin lesions assessed in this study, focusing on their characteristics and clinical aspects to establish the medical context of our research. The second section offers an overview of deep learning technologies, emphasizing the CNN general architecture and its applicability in image recognition and medical diagnostics. Lastly, the section on Skin Lesion Datasets reviews the most popular data sources used in skin lesion classification, including the strengths and limitations of these databases.

2.1 Skin Lesions

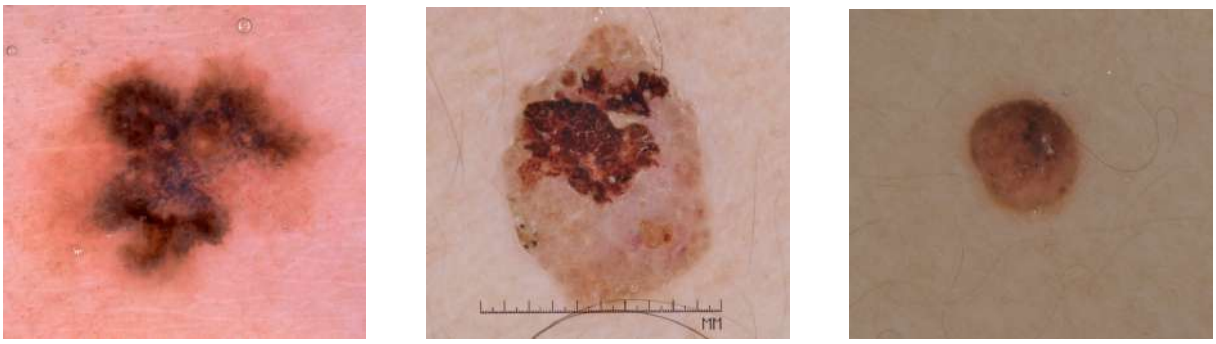
Skin lesions enclose a wide range of abnormalities or anomalies that appear on the skin surface, often due to underlying medical conditions, infections, or environmental factors [12]. These can manifest in various forms, such as bumps, lumps, sores, ulcers, or discolored areas, and can vary considerably in size, shape, texture, and color. Skin lesions can range from benign conditions, like benign nevi or seborrheic keratosis, to life-threatening diseases like melanoma, which is a form of skin cancer.

Nevi, also known as moles, are skin formations that occur when melanocytes, the

pigment-producing cells in the skin, aggregate in a specific area. The term "benign nevi" refers to moles that are non-cancerous and generally harmless, posing no significant health risks [13].

Seborrheic keratosis is a benign epidermal skin tumor characterized by its waxy, verrucous surface and a "stuck-on" appearance. These lesions are generally considered innocuous and often do not necessitate medical intervention. Nevertheless, accurate diagnosis must differentiate them from benign and malignant cutaneous neoplasms. An abrupt proliferation in the number or dimensions of these lesions may serve as an indicator of underlying systemic malignancy [14].

Melanoma is a type of skin cancer characterized by the unregulated growth of melanocytes. Despite making up a small percentage of all skin cancers, melanoma is the most deadly form of this type of cancer, and its incidence has been increasing in the world. If not diagnosed and treated early, it can spread to other parts of the body, raising a threat to the patient's life [15]. Figure 2.1 shows one example of each skin lesion class of the mentioned.



(a) (b) (c)

Figure 2.1: Examples of dermoscopic skin lesion images from ISIC 2017 dataset: (a) melanoma; (b) seborrheich kerastosis; and (c) common nevus.

In 2020, an estimated 325,000 new melanoma cases and 57,000 deaths due to melanoma were reported globally. The incidence rates were highest in Australia/New Zealand, followed by Western Europe and North America. The mortality rates were also highest in Australia/New Zealand. If the rates from 2020 remain stable, the global burden from

melanoma is projected to rise to 510,000 new cases and 96,000 deaths by 2040 as it is shown in Figure 2.2 [16].

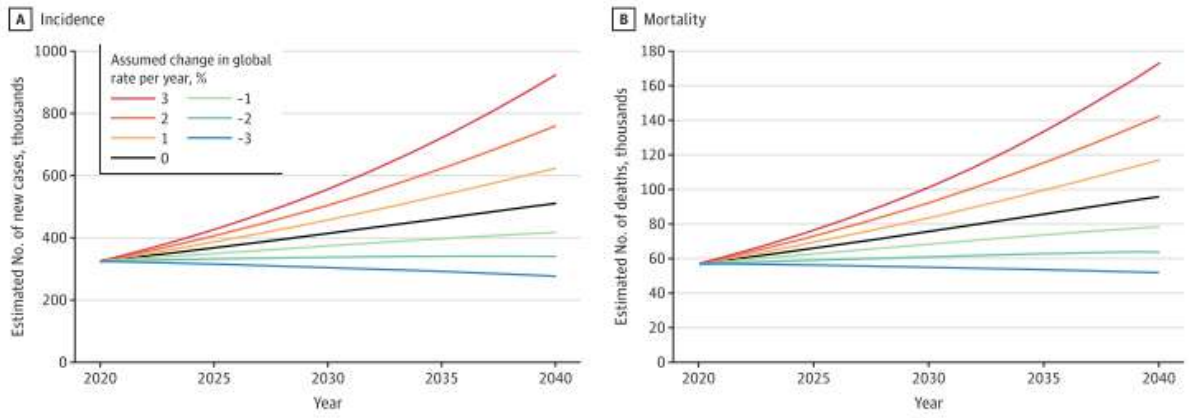
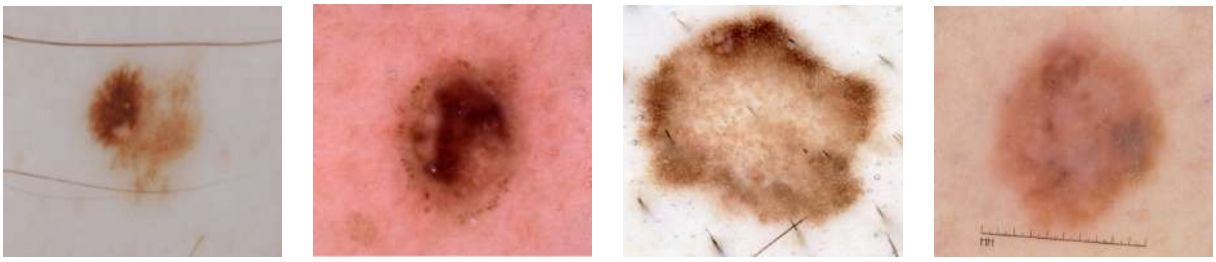


Figure 2.2: Estimated Numbers of Melanoma Cases and Deaths From 2020 to 2040, by Projection Scenario. Extracted from [16].

2.2 Clinical diagnosis approaches

Several methods are used in dermatology for the initial evaluation of skin lesions. One of the most popular methods is the ABCD criteria, which consists of evaluating four lesion features: (A) lesion asymmetry, (B) border irregularity, (C) color variation, and (D) lesion diameter greater than 6 mm. In Figure 2.3, we present examples for each of them. The ABCDE criteria add "Evolution", monitoring changes in the lesion over time [17].



(a) (b) (c) (d)

Figure 2.3: Examples of lesions presenting each of the characteristics evaluated in ABCD criteria: (a) asymmetry; (b) border irregularity; (c) color variation and (d) diameter greater than 6 mm.

The seven-point checklist consists of evaluating the presence of three main features, namely, atypical network, blue-white veil, and atypical vascular pattern, and four minor features, which are irregular dots, streaks, blotches, and regression structures. A lesion that scores 3 or more on this checklist indicates the need for further examination [18].

The Ugly Duckling Sign compares a suspicious lesion to the patient's other lesions, focusing on noticeable differences [19]. The Menzies method uses color and shape features to differentiate benign lesions from malignant ones [20]. Lastly, the Chaos and Clues approach consists of first identifying the "chaos" (asymmetry of color or structure) in a lesion, then seeking at least one of eight "clues" to malignancy [21]. All these methods serve as initial assessments and a professional medical examination is necessary for a definitive diagnosis of skin cancer.

Another non-invasive procedure often used for the early detection and diagnosis of skin cancers is dermoscopy, also known as epiluminescence microscopy. The procedure begins with the application of a liquid medium, such as alcohol or oil, onto the skin to reduce surface reflection. Then, the dermatologist places the handheld device, known as a dermatoscope, equipped with a magnifying lens and a light source, onto the skin lesion. By using this technique, it is possible to conduct a meticulous analysis of the skin structures that are not visible to the naked eye, such as pigmentation patterns and vascular structures in the epidermis and dermis. Based on these patterns and structures, the dermatologist can differentiate between benign and malignant skin lesions [22].

However, it is important to note that the diagnostic accuracy of dermoscopy significantly depends on the experience of the examiners. In 2016, Carrera et al. [23] conducted a study to analyze the reliability and validity of dermoscopic criteria in diagnosing skin lesions and identified some constraints related mostly to limitations in human judgment. They noted low levels of interobserver agreement, which can be attributed to either the lack of training of some participants or to the potential biases in diagnosis criteria selection influenced by their previous experiences with dermoscopy. Also, technical nuances, like slight variations in imaging and unstandardized viewing conditions, can impact diagnoses.

Finally, they highlight the need to improve the standardization of training methods, clarify any ambiguous criteria, and explore effective teaching methodologies in dermoscopy to improve dermoscopy's accuracy.

Diagnosing skin lesions can be a challenging task. Therefore, to confirm the diagnosis, a biopsy is often performed, wherein a sample of the skin lesion is surgically removed and analyzed in a laboratory. The biopsy can confirm the cancer's presence, identify the cells' type and depth, and guide subsequent treatment strategies. This procedure provides an accurate diagnosis; however, it is time-consuming, invasive, and sometimes painful [24].

Accurate classification and diagnosis of skin lesions are essential for determining the appropriate treatment strategy and can have a considerable impact on patient outcomes. As skin lesions can be symptomatic of various underlying conditions, their study and accurate identification are essential in dermatology and general healthcare.

2.3 Deep learning

Deep Learning (DL) is a subfield of machine learning that contemplates architectures of artificial neural networks composed of multiple layers. These neural networks try to learn from multi-layered representations of data, enabling them to understand both simple and complex features, therefore handling complex tasks in image recognition, natural language processing, and other domains requiring the interpretation of high-dimensional data [25].

A deep neural network's overall structure is composed of one input layer, a set of hidden layers with non-linear activation functions such as Rectified Linear Unit (ReLU) or Sigmoid, and one output layer, often with a SoftMax activation for classification tasks [26]. The network is trained using optimization algorithms like Stochastic Gradient Descent (SGD), Adam, or RMSprop, with the aim of minimizing a task-specific loss function, such as mean squared error for regression or cross-entropy for classification [25].

The optimization occurs through a process known as backpropagation, which computes the gradients of the loss function to update the network's weights. The goal is for the network to generalize well, making accurate predictions on new, unseen data [27].

To prevent overfitting during this training phase, regularization techniques like dropout, batch normalization, and L1/L2 regularization are often employed [28]. The learning rate is a crucial hyperparameter in these optimization algorithms, affecting both the speed and effectiveness of the training process [29].

Deep Neural Network (DNN) architectures have evolved to address various complex tasks across multiple domains. Among these, Recurrent Neural Network (RNN) are suited for sequential data and are used in natural language processing and time-series analysis [30]. Generative Adversarial Network (GAN) [31] is used for unsupervised learning, finding applications in image generation, data augmentation, and style transfer. CNN [32] is designed to process data with a grid-like topology, most commonly images. Therefore, they are used for image-related tasks such as image classification and object detection.

2.3.1 Convolutional Neural Networks

Ultimately, this work consists of classifying images into three classes. Due to the complexity of the images, the most suitable strategy is to use deep learning approaches, specifically CNNs.

The CNN architecture is built to learn spatial hierarchies of features and has as its core convolutional layers. These layers apply convolutional filters to an input image: the filters are often initialized with random values and slide across the input to produce a feature map. Stride and padding are the parameters used to control the spatial dimensions of the output feature maps. Stride refers to the step size that the filter takes as it moves across the image, while padding involves adding a specified number of zero-valued pixels around the input image border [27].

As the network goes through the convolutional layers, it gradually picks up more complex features, such as shapes and textures, based on the information extracted by the previous layers. The deep layers of the network capture abstract and high-level attributes, such as the identification of objects within the image [33].

Following the convolutional layers, non-linear activation functions are applied to introduce system non-linearity. The ReLU is the most commonly employed activation function in CNNs. Subsequent to the activation functions, pooling layers are often used to reduce the dimensionality of the feature map. Max pooling and average pooling are the most commonly used techniques for this purpose [27].

At the end of the network, fully connected layers are employed to perform the final classification task. These layers are dense, meaning that each neuron in a layer is connected to all neurons in the preceding and following layers. Finally, the output layer is designed to have as many neurons as there are classes in the classification problem, often employing a Softmax activation function for multi-class classification.

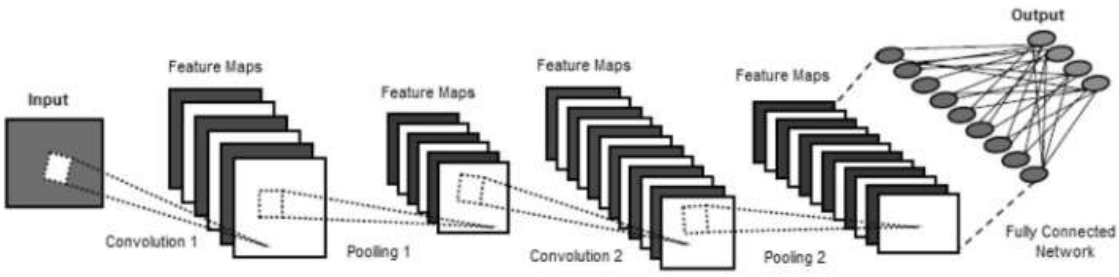


Figure 2.4: An example of the CNN architecture. Extracted from [25].

Throughout the years, multiple CNN architectures have been created, adapted, and optimized for various tasks and computational constraints [25]. One of the first models, VGGNet [34], employs a deep architecture with repetitive structures to capture increasingly complex features, setting performance benchmarks. On the other hand, ResNet [35] introduced residual connections to skip layers, enabling the training of extremely deep networks by mitigating the vanishing gradient problem. The DenseNet [36] architecture took this a step further by connecting each layer's output to every other layer in the network, thereby enhancing feature propagation and reducing the number of parameters.

EfficientNets [37] optimize performance by proportionally scaling key dimensions of the network, such as depth, width, and input resolution. This balanced approach enables higher accuracy while conserving computational resources, offering a streamlined

yet powerful architecture. Progressive Neural Architecture Search Network (PNASNet) [38] employs a machine learning-based search to optimize its architecture, ensuring a high level of efficiency and accuracy. These diverse architectures offer researchers and practitioners an array of options, each with its own advantages, complexities, and best-fit application scenarios.

2.3.2 Transfer learning

For training a deep neural network effectively, it is necessary to have a huge amount of data. For some tasks, including the classification of skin lesions, the number of data is very limited. In this case, the transfer learning technique can overcome this problem.

Transfer learning allows the use of neural networks trained for one task to be used for another task. This involves fine-tuning, which refers to performing a new round of training on the pre-trained network using another dataset.

The concept behind transfer learning is that the pre-trained models have already learned general features from an extensive dataset that serves as a great starting point. This technique usually leads to a considerable reduction in training time and computational resources, while often achieving remarkable results.

2.4 Skin Lesion Datasets

In order to develop an effective algorithm, it is crucial to train deep neural networks using extensive labeled datasets. Currently, numerous publicly accessible skin lesion datasets serve this purpose. Many of these databases are divided into three parts - training, validation, and testing - which allows for more reliable comparisons between studies that use the same dataset. This section outlines some of the most widely used skin lesion datasets and their unique characteristics.

The most popular datasets for skin lesion diagnosis are maintained by the International Skin Imaging Collaboration International Skin Imaging Collaboration (ISIC) archive, which is an open-source project that maintains some of and provides over 20000

images acquired with several devices [39]. ISIC has sponsored challenges from 2016 to 2020 that promote the use of artificial intelligence algorithms to improve lesions diagnoses accuracy [40].

The ISIC 2016 [41] dataset offers 900 dermatoscopic images for training and 379 images for testing, classified between nevi and melanoma, along with ground truth segmentation masks. ISIC 2017 [8] includes 2000, 150, and 600 dermatoscopic images for training, validation, and testing, respectively, with ground truth segmentation masks and gold standard diagnoses for melanoma, seborrheic keratosis, and benign nevi. In comparison, ISIC 2018 [39], [42] provides 10,015 training images and 193 validation images classified in 7 different diagnoses and with associated segmentation masks. Additionally, it has 1,512 test images without containing ground truth and segmentation masks.

ISIC 2019 [8], [42], [43] and ISIC 2020 [44] datasets provide 25,331 and 33,126 training images, respectively, with metadata information of gender, age, and general anatomic site, along with gold standard lesion diagnoses classification. However, both these datasets have a severe class imbalance, which can bias the trained model. ISIC 2019 and 2020 datasets do not contain any segmentation masks, and they offer additional test images without ground truth.

PH2 dataset [45], publicly available and obtained at the Dermatology Service of Hospital Pedro Hispano in Portugal, contains 200 dermatoscopic images classified into 80 common nevi, 80 atypical nevi, and 40 melanomas. For each image, a binary mask and medical annotation on diagnosis criteria are available. MEDNODE dataset contains 170 non-dermoscopic images divided into 70 melanomas and 100 nevi, obtained by the Department of Dermatology of the University Medical Center Groningen.

The DermQuest dataset [46] has 24,082 images with 134 different diagnosis classifications. However, the website for this dataset is no longer available since 2019. PAD-UFES-20 dataset [47] contains 2,298 images with six diagnoses classification and additional metadata such as the patient's age, lesion location, and other nineteen clinical data, collected using smartphones in eleven cities of Espírito Santo, Brazil.

Among these datasets, the most used in the last years for the skin lesion classification

task was the ISIC 2018, followed by the ISIC 2017 [48]. The ISIC 2018 contains more data and more class numbers, but it lacks independent test and validation sets and presents an unbalanced class sample distribution. The ISIC 2017 has fewer data and just 3 classes, but it is more balanced in between the classes and it has separated train, validation, and test sets.

Chapter 3

Literature Review

One of the primary obstacles in training CNNs for the skin lesion classification task is the lack of labeled skin lesion images available. To overcome this obstacle, transfer learning approaches are often employed. To evaluate the accuracy of the transfer-learning against trained from scratch models, Lopez et al. [49] used the VGGNet in three different ways: training from scratch, using transfer-learning and using transfer-learning and fine-tuning the architecture. The input images were pre-processed by applying pixel normalization, image cropping and resizing, and data augmentation. The better-evaluated method was the fine-tuned VGGNet with an accuracy (ACC) of 0.8133. It obtained an ACC of 0.66 by training from scratch method, and, an ACC of 0.6867 by transfer-learning method on ISIC 2016.

Hosny et al. [50] replaced the classification layer of the AlexNet with a softmax layer, and performed lesion segmentation and data augmentation. They used the DermIS-DermQuest, and MEDNODE datasets to classify into two classes and ISIC 2017 into three classes. They obtained an ACC 0.9686 for DermIS-DermQuest, 0.9770 for MEDNODE, and 0.9591 for ISIC 2017.

The exploration of different model combinations was explored by Mahbod et al. [51]. They brought together models like AlexNet, VGGNet16, ResNet-18, and ResNet-101 and used both inter and intra-network fusion. They trained a Support Vector Machine (SVM) classifier with the features extracted from these models. Their approach also

included color normalization and resizing methods. They used the ISIC 2016 and ISIC 2017 datasets for training and testing. Their approach achieved a melanoma AUC of 0.8726, seborrheic keratosys AUC of 0.9552, and the average AUC of 0.9139 on the ISIC 2017.

Hasan et al. [52] presented a method combining image preprocessing, transfer learning, and a hybrid convolutional neural network (hybrid-CNN). The initial step uses a refined DSnet to identify the lesion area, followed by data augmentation and class rebalancing. The hybrid-CNN is composed of three feature map creators, and an additional feature map is derived by merging their results. These feature maps are then classified by distinct Fully connected Layers (FC). The final outcome is the average of the outputs of the FCs. This method achieved an Area Under the ROC Curve (AUC) of 0.9600 on ISIC 2016, an AUC of 0.9500 on ISIC 2017, and an AUC of 0.9700 on ISIC 2018.

Other approaches combine sophisticated methods to identify the most important extracted features. Khan et al. [53] proposed a DenseNet with coupled Newton-Raphson-driven iteration to choose the features. Based on these features, a multilayer feed-forward neural network classifies the lesion. As preprocessing steps, contrast stretching and lesion segmentation using a faster RCNN are performed. The result for the ISIC 2016 dataset was an AUC of 0.9800, and for the ISIC 2017, an AUC of 0.9800.

Afza et al. [54] applied a ResNet-50 for categorizing and executed enhancements in contrast and segmentation of images using superpixel computation. To refine the selection of features within the network the Grasshopper algorithm is utilized. The model achieved an ACC of 0.9110 on the ISIC 2016 dataset, 0.9540 on the PH2 dataset, and 0.8580 on the HAM1000 dataset.

Additionally, other models' designs can be found in the literature. Khouloud et al. [55], after employing image pre-processing and lesion segmentation, classified the skin lesion between benign and malignant using an Inception-Resnet model. They obtained an ACC of 0.9810 on the ISIC 2016, 0.9697 on the ISIC 2017, and 0.9850 on the PH2 dataset.

Jayapriya et al. [56] employed a hybrid FCN for segmentation, and a fusion of Deep

Residual Network (DRN) with a Local texton XOR patterns (LTxXORP) is added to form a feature vector that is used as an input for the SVM to obtain the final classification (benign or malignant). The final ACC obtained is 0.8892 for ISIC 2016 and 0.8530 for ISIC 2017.

Sarkar et al. [57] used preprocessing algorithms for noise removal and image enhancement and a model based on residual learning and separable convolutional approaches to perform binary skin lesion classification. The model was trained on a subset of the ISIC archive dataset and obtained an ACC of 0.9950 for ISIC, 0.9677 on the PH2 dataset, 0.9523 on the MEDNODE dataset, and 0.9444 on the DermIS dataset.

Ge et al. [58] employed a CNN model to extract different types of features from the segmented skin lesion image. The extracted features are convolutional features, location features, statistical parameters and gray-level co-occurrence matrix features. These features are used to classify the lesions from the ISIC 2016 between malignant and benign. The model obtained an ACC of 0.92.

Skin lesion challenges, such as the ones organized by the International Skin Imaging Collaboration (ISIC), have been a boon to the field of dermatology, enabling researchers to develop and refine automated diagnostic systems that can accurately detect skin cancer. The ISIC has provided a large and publicly accessible dataset of skin images, which has been a crucial resource for advancing research in this area [59]. This dataset includes images of melanocytic and non-melanocytic skin lesions, as well as a range of benign and malignant lesions, covering a broad spectrum of skin types and conditions.

The ISIC 2017 Challenge aimed to assist participants in creating tools for image analysis, focusing on facilitating the automated diagnosis of melanoma from dermoscopic images. It was structured around three main components of skin lesion image analysis, which included the segmentation of the lesion, the identification and localization of visual dermoscopic features or patterns, and the classification of the disease. It provided approximately 2000 images as training data, around 150 images for validation, and 600 images for testing.

The winning approach, suggested by [60], implemented a system consisting of two

binary classifiers, one distinguishing melanoma from other conditions and another discerning seborrheic keratosis from the rest. This approach applied color constancy during image preprocessing and employed a ResNet model with 50 layers for the classification tasks, integrating metadata such as age and sex for enhanced accuracy.

The second place method, introduced by Iván González-Díaz [61], lesions segmentation, data augmentation, and a Structure Segmentation Network to isolate segmentation maps of specific structures important for dermatological assessments. Subsequently, a ResNet50 was utilized to categorize the lesions.

The third position was the approach proposed by Menegola et al. [62], which used lesion segmentation and ResNet101 and Inception-v4 models for the classification.

Both second and third place used the metadata provided and all the mentioned strategies incorporated data augmentation and initiated their models with pre-training on the ImageNet dataset, followed by fine-tuning on the ISIC 2017 dataset.

The challenge set by ISIC in 2019 involved classifying lesions into one of nine distinct diagnostic groups. The training set comprised over 25,000 images of eight categories, and the test dataset included an additional outlier class. Two of the leaderboard solutions are detailed below.

Gessert et al. [63] introduced a method incorporating EfficientNets, SENet154, and ResNext models. Their approach included comprehensive data augmentation and image preprocessing, incorporating strategies for cropping and resizing, as well as color constancy methods and resizing images. The final classification was given the average of the predictions from each model and the models identified through a search strategy, aiming to find the optimal subset of configurations.

Pacheco et al. [64] proposed the predictions ensemble of SENet, PNASNet, InceptionV4, ResNet-50/101/152, DenseNet-121/169/201, MobileNetV2, GoogleNet, and VGG-16/19 models. They also applied the color constancy method, and extensive data augmentation was applied.

For the ISIC 2020 Challenge, a dataset comprising over 33,000 training images and around 10,000 test images, categorized as malignant or benign, was provided [44]. In one

of the winning approaches, Ha et al. [65] introduced a method involving EfficientNet-B3 to EfficientNet-B7, SE-ResNeXt-101, and ResNeSt-101 and data augmentation techniques. The final result was the average of the 18 models‘ predictions.

Based on the recent literature, ensemble techniques and image preprocessing have shown great results in addressing the challenges in skin lesion classification. Various approaches in the ISIC Challenges use ensemble techniques, which enhances the robustness and reliability of the classification systems, allowing for the incorporation of diverse perspectives and methodologies in analyzing skin lesions.

Furthermore, image preprocessing, including techniques like lesion segmentation, color constancy, cropping, and resizing, can refine the input data to ensure the models are trained on high-quality, standardized images. This step can mitigate the impact of variations in the input data, thereby enhancing the accuracy of the subsequent classification. The consistent application of image preprocessing and data augmentation across various methods emphasizes their importance in optimizing the performance of the models.

Inspired by these methods, especially the recent ISIC Challenges, we chose 13 models, three ensemble approaches, and four image preprocessing techniques. Our goal is to evaluate how different combinations of these three factors affect the final classification of the skin lesions.

Chapter 4

Proposed Approach

Our research aims to utilize transfer learning to train multiple models on the ISIC 2017 dataset for classifying three skin lesion categories. We have also tested a variety of preprocessing techniques to assess their impact on the models. Additionally, we have evaluated the influence of data augmentation and tested three different ensemble techniques to determine the one that yields the best performance. In this section, we will provide an overview of the dataset used, the tools employed to conduct this study, and a detailed explanation of each image preprocessing technique employed in preparing the data for analysis. We will also discuss the transformations used in the data augmentation process, the transfer learning technique, the characteristics of each model, and the optimizer used during training. Additionally, we will provide insights into each ensemble technique selected, the evaluation metric employed, and the training process.

4.1 Dataset

In this work, we chose the ISIC 2017 dataset. The ISIC 2017 dataset is the second most used dataset for research related to the skin lesion classification task only staying behind the ISIC 2018 dataset [48]. The unique aspect of this dataset is its clear partitioning into three subsets: train, validation, and test. This clear separation creates a fair and balanced environment for comparing different models that use the same dataset and avoids

the necessity of time-consuming cross-validation to describe the results. By keeping the quality of data consistent across all models, the only variable becomes the training process and the model itself, allowing for a fair comparison.

The ISIC 2017 dataset comprises dermoscopic images of skin lesions categorized into three classes: melanoma (374 for training, 30 for validation, and 117 for testing), seborrheic keratosis (254 for training, 42 for validation, and 90 for testing), and benign nevi (1372 for training, 78 for validation, and 393 for testing). The dataset contains a total of 2,000 images for training, 150 for validation, and 600 for testing. In addition, the patient's gender and age were also provided for each image, but it was not used in the scope of this study.

4.2 Image preprocessing

The use of image preprocessing techniques can enhance image quality, extract relevant information, and mitigate variations in illumination and contrast, resulting in improved accuracy and generalization of deep learning models [66]. In this work, we explored the use of four different image preprocessing techniques that produce distinct effects on the images.

For the purpose of ensuring that the pre-trained network receives input data in a format that is consistent with the original training of the models on the ImageNet dataset, we have consistently employed a normalization technique by default. Furthermore, we have systematically applied image resizing to all images to make sure that their height and width conform to the input dimensions required by the pre-trained model being used. This step has been repeatedly implemented across all experimental iterations to ensure that the image data is compatible with the model's processing and analysis requirements.

One of the image processing techniques used was centered square cropping. This technique was chosen for three main reasons: first, to remove unnecessary information from the image, such as excess skin and black corners; second, to avoid distortion of the

lesion when resizing it to be introduced in the model's input while preserving the original format of the lesion; and third, because almost all images are located in the center of the image. As an example, after this procedure, an image with 600x400 dimensions would be cropped to 400x400.

The contrast enhancement technique has been used in various studies on the skin lesion classification task to aid in the feature extraction [54], [57]. We use a technique called Contrast Limited Adaptive Histogram Equalization (CLAHE) to address this in our work. This method is a variation of the Histogram Equalization (HE) technique, where the image or volume is divided into smaller blocks. The HE is then applied to each block independently, allowing for local contrast enhancement. This approach adapts the amplification of contrast to the characteristics of each block, preventing over-enhancement or amplification [67].

Skin lesion images can have artifacts such as hairs and marks, that can negatively affect the performance of the model. To address this issue, we utilized the DullRazor algorithm. This algorithm works by identifying dark regions corresponding to hair using a generalized grayscale morphological closing operation. It then checks for the characteristic thin and long structure of hair pixels and replaces the verified pixels with bilinear interpolation. Finally, the replaced hair pixels are smoothed by an adaptive median filter [68].

The practical application of all the image preprocessing techniques mentioned is presented in Figure 4.1.

(a) (b) (c) (d) (e)

Figure 4.1: Examples of preprocessing techniques applied to an image: (a) original image; (b) center crop; (c) color constancy; (d) hair removal; and (e) CLAHE.

4.3 Data Augmentation

Data augmentation is a technique employed to overcome the problem of the lack of data by artificially increasing the size of a training dataset. This is achieved through methods like data warping, where existing images are transformed while preserving their labels. Data warping includes techniques such as geometric and color transformations, random erasing, and other image transformations. This is used to improve the diversity and variability of the training data, leading to enhanced model performance and generalization. [69].

In this study, we experimented with different data augmentation processes used in previous studies with similar tasks and found that the approach employed by Pacheco et al. [64] yielded the best results. The data augmentation process in our study is described as follows:

- Zoom in: Randomly scale the image along the x and y axes with a probability of 25%.
- Horizontal flip: Horizontally flip 50% of all images.
- Vertical flip: Vertically flip 20% of all images with a probability of 20%.
- Rotate: Rotate the image by up to 120 degrees in either direction with a probability of 25%.
- Gaussian blur: Apply Gaussian blur to the image with a sigma value ranging from 0 to 1.5 with a probability of 25%.
- Dropout and Coarse Dropout: Randomly remove rectangular regions of up to 5% of the image or randomly remove up to 5% of the pixels from the image with a probability of 10%.

- Brightness, Hue and Saturation: Randomly change the image's brightness by -15 to 15 of the original value or change the hue and saturation of the image with a probability of 25%.

4.4 Transfer Learning

Transfer learning is a technique used in machine learning when there is a limited amount of data available for training. It involves selecting a pre-trained model that has already been trained with a large amount of data for a similar task and then training it again for a different task. This approach saves time and resources as we don't need to train the model from scratch. Instead, we can start with the pre-trained model's weights, which serve as a solid foundation. In other words, transfer learning allows us to use the knowledge learned from a previous task and apply it to a new one [70].

In our work, we used the PyTorch Image Models (TIMM) [71] library to download the pre-trained models previously trained in the ImageNet dataset. We prepared these models for our task by replacing the FC layer with a new one with the same number of neurons as the number of classes in our dataset. Then, we performed another training, updating all the weights in the network using the ISIC 2017 dataset.

4.4.1 Pre-trained models

Based on what was used in the recent literature, we chose 13 models of 7 different architectures to be evaluated in our work: EfficientNet (B0-B6), PNASNet5, ResNeSt101, ResNeXt101, SEResNeXt101, DenseNet121, VGG19.

EfficientNet [72] is a group of CNNs that utilize a scaling approach to balance the network's depth, width, and resolution for optimal performance. This is achieved through a compound scaling method that uniformly scales these dimensions using a compound coefficient. The scaling levels - B0, B1, B2, and so on - represent different base model

versions. Each version is progressively larger, potentially providing higher accuracy and requiring more computational resources.

The Progressive Neural Architecture Search (PNASNet) [73] is a model that utilizes a sequential model-based optimization strategy to identify the best neural network architecture for a given task. This approach entails a search strategy that progressively increases the complexity of the model by adding more building blocks, starting with simple cells and gradually increasing in complexity. For this work, we used the PNASNet-5Large model that achieved state-of-the-art performance on the ImageNet dataset. This was constructed by stacking the optimal cell structures discovered during the PNAS search process. The "5" refers to the number of times the normal cell is repeated in the architecture, and "Large" refers to the fact that this is a more complex variant than other potential architectures derived from the PNAS approach.

The ResNeSt101 [74] is a variant of the ResNet model that includes the “Split-Attention" block. This feature allows attention to be distributed across different groups of feature maps. Our work utilized the ResNeSt101, which consists of 101 layers. The ResNeSt architecture maintains the residual connections that help mitigate the vanishing gradient problem, allowing for the training of very deep networks.

The ResNeXt101 [75] is a variation of the ResNet model that introduces a new dimension called “cardinality". This dimension refers to the number of independent paths in the network, which helps to increase the model’s capacity and performance without significantly increasing computational cost. Like the ResNeSt, the architecture of ResNeXt maintains the residual connections of ResNet. Our work used the ResNeXt101, which consists of 101 layers.

The SEResNeXt [76] is a variant of the ResNeXt model that incorporates the Squeeze-and-Excitation (SE) blocks. These blocks adaptively recalibrate the feature responses of each channel by modeling the interdependencies between them. This enables the model to emphasize informative features and suppress less useful ones, enhancing the representational power of the network. Our work used the SEResNext101 model, which consists of 101 layers.

The Dense Convolutional Network (DenseNet) [77] is a type of CNN where each layer is connected to all other deeper layers in the network. Unlike traditional CNNs, in DenseNet, each layer acquires the feature maps of all previous layers as inputs, and its feature maps are used as inputs into all successive layers. This dense connectivity pattern facilitates feature reuse throughout the network, significantly reduces the number of parameters, and improves the flow of information and gradients throughout the network, making it easier to train. In our work, we used a DenseNet with 121 layers.

Finally, the Visual Geometry Group (VGG) model [34] is a CNN characterized by its architecture uniformity. All hidden layers are convolutional, using a minimal receptive field. The convolution stride remains fixed at 1 pixel to maintain spatial resolution after convolution. The network follows this with a max-pooling step with a 2x2 pixel window and stride 2. In our work, we used the VGG model composed of 19 layers.

4.5 Ensemble techniques

In recent years, ensemble techniques have been found to be effective in addressing the challenges of skin lesion classification and have been explored in several approaches for this task [63]–[65]. This approach is able to improve the classification systems' robustness and reliability by gathering the strengths of each model [78].

In our study, we analyzed the outcomes of three different ensemble techniques. The first technique is the "Average of all" ensemble method, which aggregates predictions from all individual models, irrespective of their individual performance or preprocessing configurations. It calculates the average prediction probability for each class across the entire ensemble. This approach ensures that no single model's predictions dominate the final outcome, resulting in a balanced and unbiased combination of all models.

The second ensemble technique is the "Average of 3". This technique adopts a more selective approach. It identifies the best combination of 3 models from all possible model combinations, considering the entire ensemble. This method leverages the strengths of

these selected model combinations while disregarding weaker performers.

Finally, the "Voting" ensemble operates differently by aggregating predictions through a majority voting mechanism. It assigns a class label based on the most commonly predicted class among the ensemble of models. This method can be particularly effective when models have diverse decision boundaries, as it helps mitigate errors caused by individual model biases.

4.6 Training Process

For the training process, we used a GitHub named RAUG [79] that offers a pipeline to train deep learning models using PyTorch. We used the Cross-Entropy loss function in all of our models, as it is widely used in classification tasks. To tackle the issue of class imbalance in the dataset, we incorporated weights for different classes in the loss function. This allowed the model to give equal importance to both majority and minority classes, thus making it more sensitive to the minority class and addressing the class imbalance.

To optimize the performance of each model, we selected the hyperparameters such as optimizer, initial learning rate, and batch size through a series of experiments. In addition, we used a learning rate scheduler to dynamically adjust the learning rate during training based on the observed plateauing of the validation loss. This approach aims to improve the overall performance and convergence of the model.

Specifically, we set the learning rate scheduler with a patience number of 10 epochs, a minimum learning rate of 1×10^{-6}, and a reduction factor of 0.1. If the loss does not improve within ten epochs, the learning rate will be reduced by a factor of 0.1 until it reaches the minimum threshold of 1×10^{-6}. Through this approach, we ensure that the model's learning rate is appropriately adjusted during training to facilitate better convergence and performance.

The purpose of setting the patience number to 10 epochs is to allow the model to learn and improve for a certain amount of time before reducing the learning rate. This helps to avoid overfitting and ensures that the model is given enough time to improve

before the learning rate is decreased. By setting a minimum learning rate of 1×10^{-6}, we avoid reducing the learning rate to an excessively low value that may hinder the model's learning ability. Through these hyperparameter settings and learning rate scheduler, we aimed to optimize the performance of each model and improve its convergence.

In addition, we have implemented an early stopping mechanism in our training process. This mechanism terminates the training process if the balanced accuracy does not improve for a specific number of epochs. The number of epochs for early stopping varies depending on the experiment. Furthermore, we save the best epoch model based on balanced accuracy to avoid overfitting and obtain optimal performance of the model.

4.7 Evaluation Metrics

Balanced accuracy was the primary metric used in our work to rank the different approaches we tried. It measures the average accuracy of each class, weighted by the number of samples in each class. This metric is useful when the classes have a different prevalence in the dataset and is defined as (TPR + TNR) / 2, where TPR is the true positive rate (also known as recall or sensitivity), and TNR is the true negative rate (the fraction of true negatives among the total number of actual negatives). It is described by the equation 4.1.

$$\frac{TPR + TNR}{2} \tag{4.1}$$

Another metric used to compare with other works was the Area Under the ROC Curve (AUC). The ROC curve is a plot of the TPR against the False Positive Rate (FPR) at various classification thresholds. AUC is a single-number summary of the ROC curve that measures the classifier's ability to distinguish between the positive and negative classes. AUC ranges from 0 to 1, with a higher value indicating better performance. AUC is a helpful metric for evaluating skin lesion classification models because it is less sensitive to imbalanced datasets than other metrics like accuracy or precision.

Chapter 5

Experimental Results and Analysis

5.1 Results

The first step to achieve the optimal model configuration involved determining the most suitable hyperparameters for each model. To accomplish this, we conducted a series of experiments utilizing the SGD, Adam, and AdamW optimizers, employing initial learning rates of 1×10^{-3} and 1×10^{-4}. Furthermore, we tested batch sizes ranging from 4 to 64, considering the GPU constraints for each model. We trained each model for 100 epochs, implementing an early stopping criterion after 15 epochs. Finally, we evaluated the optimal hyperparameters for every model by analyzing the balanced accuracy metrics on the test data. The final hyperparameters configuration for each model is presented in Table **??**.

The next step was to evaluate the performance of each model when employing preprocessing steps to the data. For better understanding, each preprocessing algorithm is associated with a letter as follows:

- A: Contrast Enhancement
- B: Color Constancy
- C: Hair Removal

Model	Batchsize	Initial learning rate	Optimizer
EfficientNet-B0	8	0.001	AdamW
EfficientNet-B1	8	0.001	AdamW
EfficientNet-B2	8	0.001	Adam
EfficientNet-B3	32	0.001	Adam
EfficientNet-B4	16	0.001	AdamW
EfficientNet-B5	8	0.0001	AdamW
EfficientNet-B6	4	0.0001	AdamW
ResNeSt101	8	0.0001	AdamW
SeResNeXt101	32	0.0001	Adam
VGG19	8	0.001	SGD
DenseNet121	32	0.0001	Adam
ResNeXt101	8	0.001	SGD
PNASNet5	8	0.0001	Adam

- D: Center Crop

In Table 5.1, the balanced accuracy of each experiment obtained by the combination of different preprocessing (PP) techniques is presented. The preprocessing combinations that performed better than the raw data are shown in bold and the best result for each model is highlighted with a text box.

Table 5.1: Image Preprocessing: Experiments Results

Model	Raw data	A	B	C	D	AB	ABC	ABCD	AC	ACD	AD	BC	BCD	BD	CD
DenseNet121	0.6753	0.6716	**0.6923**	**0.7176**	**0.6974**	**0.6760**	0.6092	0.6330	0.6421	0.6382	0.6705	0.6282	0.6534	**0.7000**	0.6574
EfficientNet-B0	**0.7073**	0.6871	0.6529	0.6960	0.6623	0.6573	0.6304	0.6096	0.6149	0.6408	0.6509	0.6968	0.6645	0.6388	0.6583
EfficientNet-B1	0.6654	**0.6716**	**0.6840**	0.6601	**0.6880**	0.6336	0.6440	**0.6944**	0.6412	0.6558	**0.6966**	**0.7016**	0.6255	**0.6801**	0.6623
EfficientNet-B2	0.6993	0.6957	0.6445	**0.7091**	0.6572	0.6542	0.6600	0.6262	0.6156	0.6129	0.6611	0.6833	0.6928	0.6372	0.6806
EfficientNet-B3	0.6767	**0.7120**	**0.6982**	**0.7248**	**0.6830**	**0.7175**	0.6612	**0.7260**	**0.6866**	**0.7046**	**0.6881**	**0.7010**	**0.7278**	**0.7286**	**0.7195**
EfficientNet-B4	0.7216	0.6944	0.7039	0.7093	0.6734	0.6547	0.6493	0.6470	0.6983	0.6490	0.7029	**0.7317**	**0.7284**	**0.7364**	0.7042
EfficientNet-B5	0.6976	**0.7113**	0.6522	**0.7099**	**0.7250**	**0.7163**	0.6814	0.6958	**0.7130**	**0.7093**	**0.7152**	0.6816	0.6846	**0.7340**	0.6841
EfficientNet-B6	0.6874	**0.6943**	**0.7395**	**0.7213**	0.6735	**0.6943**	**0.7057**	0.6856	0.6599	0.6806	**0.7113**	**0.7139**	**0.6953**	**0.7383**	0.6871
PNASNet	0.7406	0.6718	0.7271	0.7398	**0.7522**	0.7175	0.6964	0.6783	0.7051	0.7015	0.7127	0.7123	0.6871	0.7406	0.7147
ResNeSt101	0.7344	0.7104	**0.7620**	**0.7447**	0.6993	0.7017	0.6706	0.7306	0.6842	0.6768	**0.7447**	0.7083	0.7070	0.6988	0.7286
ResNeXt101	0.6828	0.6731	**0.7281**	**0.6957**	**0.7194**	**0.7148**	0.6535	**0.6950**	**0.6914**	0.6787	**0.7104**	**0.7253**	**0.6967**	**0.6854**	**0.6893**
SEResNeXt101	0.7310	0.7032	**0.7481**	0.6863	0.7117	0.6887	0.6650	0.6634	0.6491	0.7065	0.6990	0.7165	0.6939	0.7184	0.6740
VGG19	**0.6922**	0.6011	0.6450	0.6606	0.6121	0.6626	0.6546	0.6532	0.6512	0.6504	0.6662	0.6506	0.6142	0.6337	0.6103

Analyzing the performance of each model, we notice that models such as EfficientNet-B0, EfficientNet-B4, SeResNeXt101, ResNeSt101, and PNASNet demonstrate balanced accuracy values exceeding 0.7, indicating their potential effectiveness in this classification task. Furthermore, PNASNet5 achieves the highest balanced accuracy score of 0.7406,

suggesting its superior performance. On the other hand, models such as EfficientNet-B1, EfficientNet-B2, EfficientNet-B3, EfficientNet-B5, EfficientNet-B6, VGG19, DenseNet121, and ResNeXt101 exhibit relatively lower balanced accuracy values below 0.7, indicating the need for further improvement. These results are shown in Figure 5.1.

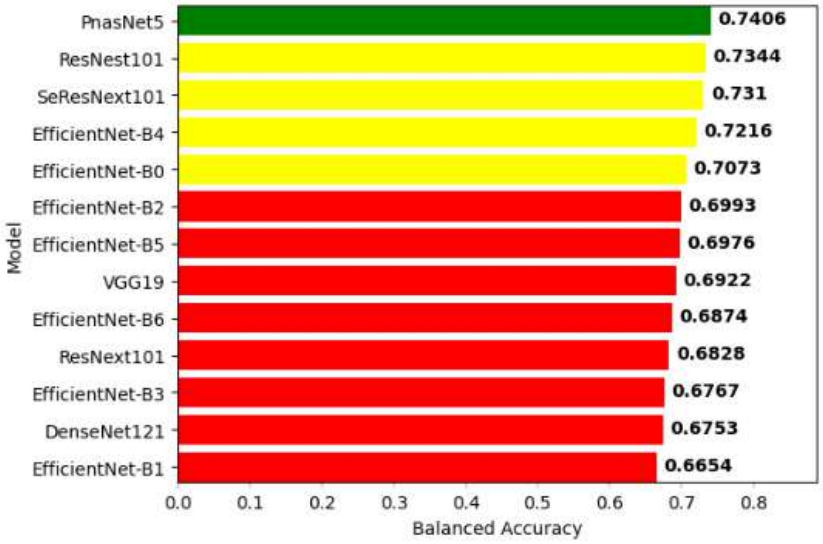


Figure 5.1: Balanced accuracy of all models before PP and DA.

Analyzing the results of the image preprocessing experiments performed on each model presented in Table 5.1, we observed an overall improvement in model performance with preprocessing, except for EfficientNet-B0 and VGG19. Additionally, PNASNet5Large and SEResNeXt101 showed improvement for only one experiment. Among all the other methods, the application of Color Constancy, Hair Removal, and the combination of Color Constancy and Center Crop improved the performance of 7 models each, with the highest number of improvements.

Despite being the best model when trained on raw data, PNASNet5Large does not improve its performance with most image preprocessing techniques, unlike other models. One hypothesis to this is that the pre-trained model's architecture may already be able to recognize a wide range of features, making it sensitive to changes introduced by certain preprocessing techniques.

The model that showed the most significant improvement was the EfficientNet-B6,

which increased its balanced accuracy by 0.0520 with the application of Color Constancy. However, the ResNeSt101 had the highest score among all these experiments, achieving a balanced accuracy of 0.7620 with the application of Color Constancy only. Nevertheless, the experiments that included Contrast Enhancement did not outperform the other configurations.

The second set of experiments involved the addition of the data augmentation process. To optimize the performance of the model, a number of adjustments were made to the training process. One of these adjustments was an increase in the maximum number of epochs to 150 since we noticed that the model was not fully converged after the initial 100 epochs, and these additional epochs allowed the model to learn and improve its performance. In addition, the initial learning rate was decreased to 1×10^{-4}. This adjustment was made in order to help the model gradually adjust the weights of the model during training, leading to a more stable convergence of the training process. Overall, these adjustments were made with the aim of giving the model more opportunities to learn and converge to a better solution, thereby optimizing its performance on the given task. Table 5.2 presents the balanced accuracy of the data augmentation for each model associated with different image preprocessing techniques.

Table 5.2: Data augmentation: Experiments Results

Model	Raw data	A	B	C	D	AB	ABC	ABCD	ABD	AC	ACD	AD	BC	BCD	BD	CD
DenseNet121	0.7182	**0.7216**	**0.7450**	0.7120	0.7002	**0.7392**	0.6560	0.6765	0.7014	0.6817	0.6850	**0.7492**	**0.7239**	0.7119	0.7025	**0.7471**
EfficientNet-B0	0.7170	0.6788	**0.7182**	**0.7176**	**0.7343**	0.6912	0.6656	**0.7176**	0.6969	0.6881	**0.7250**	**0.7195**	0.7020	**0.7207**	**0.7181**	**0.7199**
EfficientNet-B1	0.7064	0.6775	**0.7094**	**0.7088**	0.6974	0.6963	0.6635	0.7026	**0.7259**	0.6711	0.6866	**0.7401**	**0.7427**	0.6976	**0.7194**	**0.7156**
EfficientNet-B2	0.7006	**0.7054**	**0.7266**	**0.7382**	**0.7440**	**0.7387**	0.6620	0.6791	**0.7617**	0.6729	0.6795	**0.7405**	**0.7116**	**0.7435**	**0.7266**	**0.7156**
EfficientNet-B3	0.7356	**0.7370**	**0.7427**	**0.7657**	**0.7634**	0.7320	0.6958	0.7288	**0.7529**	0.7000	0.7237	**0.7384**	**0.7481**	**0.7421**	**0.7526**	**0.7888**
EfficientNet-B4	0.7441	0.7363	**0.7538**	0.7310	**0.7458**	**0.7449**	0.6775	0.6990	**0.7550**	0.6876	0.7060	0.7360	**0.7654**	0.7408	0.7441	**0.7452**
EfficientNet-B5	0.7653	0.6991	0.7378	0.7229	**0.7718**	0.7406	0.7498	0.7428	0.7335	0.7321	0.7355	0.7550	0.7429	**0.7785**	**0.7896**	0.7506
EfficientNet-B6	0.7651	0.7348	0.7549	0.7387	0.7478	0.7329	0.7282	0.7606	0.7589	0.7631	0.7383	0.7514	0.7174	**0.7686**	0.7217	0.7594
PNASNet	0.7412	0.7274	**0.7733**	0.7322	**0.7597**	**0.7583**	0.7251	0.7034	0.7241	0.6806	0.6997	**0.7487**	**0.7551**	**0.7690**	**0.7563**	**0.7487**
ResNeSt101	0.7459	**0.7513**	0.7414	0.7168	**0.7566**	0.7167	0.7154	0.7036	**0.7527**	**0.7615**	0.7099	0.7313	**0.7555**	**0.7560**	**0.7730**	0.7252
ResNeXt101	0.7016	**0.7191**	**0.7267**	0.6948	0.6829	**0.7071**	0.6671	0.6734	**0.7179**	0.6702	**0.7054**	0.6909	**0.7330**	**0.7091**	0.7001	**0.7089**
SEResNeXt101	0.7432	0.7299	0.6931	0.7354	0.7251	0.7201	0.6866	0.7170	0.7125	0.7068	0.7017	**0.7503**	0.7086	0.7181	0.7062	0.7108
VGG19	0.6786	0.6783	**0.6951**	0.6783	**0.7142**	**0.6800**	**0.6840**	0.6386	**0.7040**	0.6639	0.6716	**0.7192**	**0.7053**	0.6633	**0.7108**	**0.6996**

Overall, the use of data augmentation techniques improved model performance. Figure 5.2 presents a chart comparing the performance of the models trained with and without data augmentation and both without applying preprocessing techniques. We noticed

that the application of data augmentation improved the model's performance. The only exception is the VGG19 model, which decreased its balanced accuracy in 0.0136. For EfficientNet-B5 and EfficientNet-B6, the performance improvement due to data augmentation was significant. This indicates that data augmentation can benefit larger and more complex models.

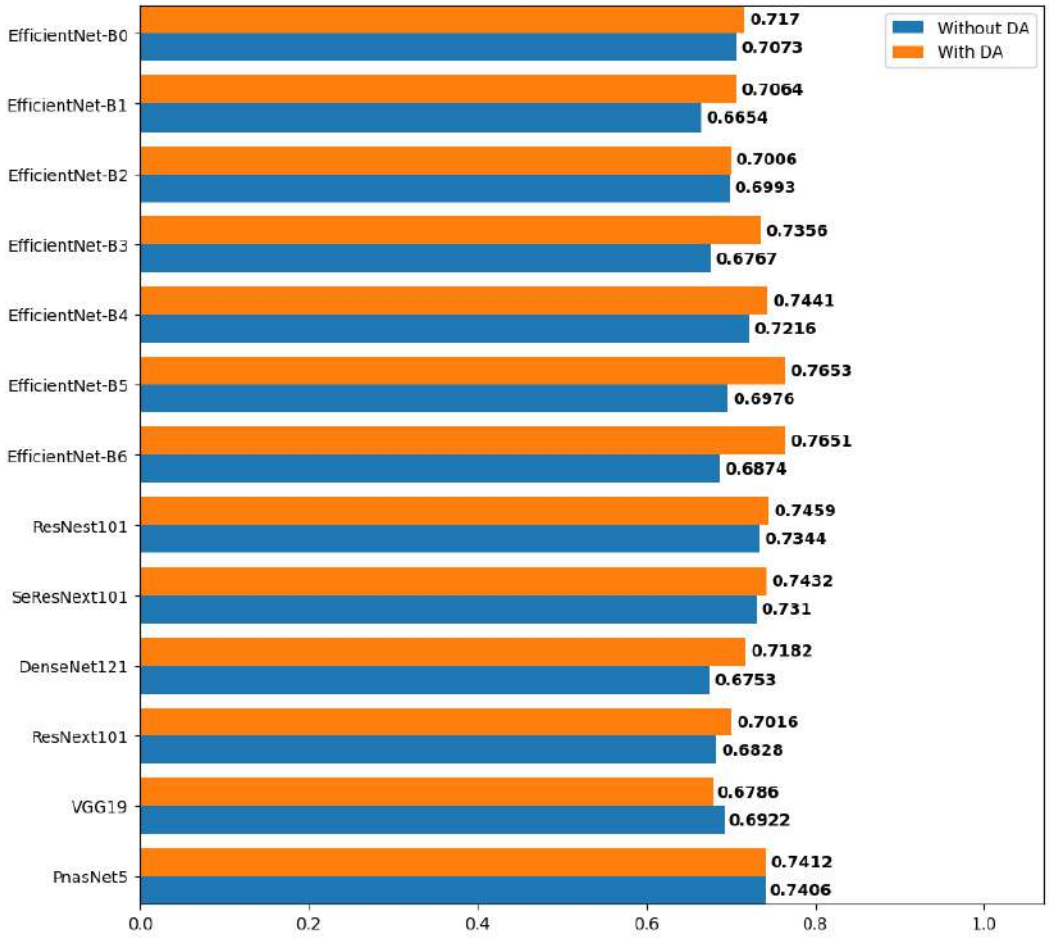


Figure 5.2: Data augmentation results for each model.

By analyzing the results of the image preprocessing experiments performed on each model presented in Table 5.2. We observed the combination of preprocessing and data augmentation improved the model performance compared to data augmentation only. Among various methods employed, the application of data augmentation combined with Color Constancy, the fusion of Color Constancy and Hair Removal, and the combination

of Hair Removal and Center Crop were found to enhance the performance of 9 models significantly. These three methods produced the highest number of improvements.

Among the various models tested, the most significant improvement was observed in the case of EfficientNet-B2, which increased its balanced accuracy by 0.0611 with the application of Contrast Enhancement, Color Constancy, and Center Crop. However, the highest score was achieved by EfficientNet-B5 with a balanced accuracy of 0.7896 when Color Constancy and Center Crop were applied. Two models, EfficientNet-B6 and SEResNeXt101, showed low sensitivity to applying preprocessing techniques with data augmentation. They only showed a minor improvement for one method. Contrary to the results of experiments without data augmentation, combining Contrast Enhancement with other methods significantly improved the four models' performance.

It's worth noting that while the EfficientNet-B6 model exhibited the greatest improvement when subjected to the preprocessing technique alone and also had the highest performance and greatest performance increase with data augmentation alone, combining both techniques led to a decline in performance in most experiments. Although one experiment showed an improvement for this model, it was not significant. In contrast, the EfficientNet-B5 improved its performance with data augmentation only and achieved the best performance overall models with the combination of both techniques.

Next, ensemble techniques were applied for each experiment. For the image preprocessing experiments without data augmentation, we obtained the results presented in Table 5.3. Each row represents the balanced accuracy score of the ensemble techniques for each preprocessing configuration. The best result for each model is highlighted in red text, and the preprocessing combinations that outperformed the raw data are in blue, and Table 5.4 presents the models used in each experiment of the "Average of 3" ensembles.

Table 5.3: Image Preprocessing: Ensemble Results

Ensemble	Raw data	A	B	C	D	AB	ABC	ABCD	AC	ACD	AD	BC	BCD	BD	CD
Average of all	0.7547	**0.7674**	**0.7685**	**0.7713**	**0.7690**	**0.7596**	0.7146	**0.7657**	**0.7599**	0.7500	**0.7807**	**0.7769**	**0.7570**	**0.7613**	**0.7693**
Average of 3	0.7671	0.7620	**0.7846**	**0.7879**	**0.7703**	0.7634	0.7565	**0.7808**	0.7580	0.7669	**0.7828**	**0.7798**	**0.7697**	**0.7767**	**0.7760**
Voting	0.7371	**0.7617**	**0.7718**	**0.7693**	**0.7647**	**0.7451**	0.7166	**0.7591**	**0.7534**	**0.7491**	**0.7705**	**0.7678**	**0.7493**	**0.7621**	**0.7624**

In the ensemble results presented in Table 5.3, we observe that the highest score

Table 5.4: Image Preprocessing: Average of 3 ensemble results

Experiment	Models	Balanced accuracy
Raw data	EfficientNet-B0, ResNeSt101, SEResNeXt101	0.7671
A	EfficientNet-B3, EfficientNet-B6, ResNeSt101	0.7620
B	EfficientNet-B3, EfficientNet-B6, ResNeXt101	0.7846
C	tEfficientNet-B3, EfficientNet-B5, PNASNe	**0.7879**
D	EfficientNet-B5, PNASNet, ResNeSt101	0.7703
AB	EfficientNet-B3, EfficientNet-B5, PNASNet5Large	0.7634
ABC	PNASNet, ResNeSt101, ResNeXt101	0.7565
ABCD	EfficientNet-B5, ResNeSt101, VGG19	0.7808
AC	EfficientNet-B1, EfficientNet-B5, ResNeSt101	0.7580
ACD	EfficientNet-B3, EfficientNet-B6, SEResNeXt101	0.7669
AD	EfficientNet-B1, EfficientNet-B5, ResNeXt101	0.7828
BC	EfficientNet-B3, EfficientNet-B4, ResNeXt101	0.7798
BCD	EfficientNet-B2, EfficientNet-B4, PNASNet5Large	0.7697
BD	EfficientNet-B6, PNASNet, SEResNeXt101	0.7767
CD	EfficientNet-B3, EfficientNet-B5, ResNeSt101	0.7760

achieved was for the "Average of 3" ensemble with the C configuration, corresponding to the employment of the Hair Removal. Next, the "Average of all" ensemble with the application of Contrast Enhancement and Center Crop, followed by the "Voting" ensemble with Color Constancy.

We also tested the ensemble techniques for the experiments with data augmentation. Table 5.5 presents the result for each ensemble technique given the combination of different preprocessing techniques, and Table 5.6 presents the models used in each experiment of the "Average of 3" ensembles.

Table 5.5: Data augmentation: Ensemble Results

Ensemble	Raw data	A	B	C	D	AB	ABC	ABCD	ABD	AC	ACD	AD	BC	BCD	BD	CD
Average of all	0.7811	**0.7904**	**0.7818**	0.7779	**0.8007**	**0.7830**	0.7529	**0.7845**	**0.7907**	0.7757	**0.7836**	**0.7989**	**0.7830**	**0.8064**	**0.7904**	**0.8020**
Average of 3	0.7939	0.7900	**0.8044**	**0.7977**	**0.8058**	0.7876	0.7720	**0.8036**	**0.8130**	**0.7964**	0.7914	**0.8047**	**0.7989**	**0.8132**	**0.8100**	**0.8106**
Voting	0.7882	**0.7890**	0.7744	0.7747	**0.7987**	0.7864	0.7537	0.7714	0.7865	0.7555	0.7824	**0.8029**	0.7785	**0.7910**	0.7719	**0.8000**

For the ensemble results exploiting data augmentation associated with preprocessing techniques presented in Table 5.5, we observe that the highest score achieved was for the "Average of 3" ensemble with the BCD configuration, corresponding to the employment of Color Constancy, Hair Removal, and Center Crop. Next, the "Average of all" ensemble with the same configuration, and finally, the "Voting" ensemble with AD configuration corresponding to the combination of Contrast Enhancement and Center Crop.

Table 5.6: Data augmentation: Average of 3 ensemble results

Experiment	**Models**	**Balanced accuracy**
Raw data	EfficientNet-B3, EfficientNet-B6, ResNeSt101	0.7939
A	DenseNet121, ResNeSt101, ResNeXt101	0.7900
B	EfficientNet-B3, EfficientNet-B6, PNASNet5Large	0.8044
C	EfficientNet-B1, EfficientNet-B2, EfficientNet-B3	0.7977
D	EfficientNet-B5, EfficientNet-B6, VGG19	0.8058
AB	DenseNet121, EfficientNet-B4, PNASNet5Large	0.7876
ABC	EfficientNet-B5, EfficientNet-B6, VGG19	0.7720
ABCD	EfficientNet-B0, EfficientNet-B5, SEResNeXt101	0.8036
ABD	EfficientNet-B2, EfficientNet-B3, ResNeSt101	0.8130
AC	DenseNet121, EfficientNet-B6, ResNeSt101	0.7964
ACD	EfficientNet-B3, EfficientNet-B6, VGG19	0.7914
AD	EfficientNet-B0, EfficientNet-B6, SEResNeXt101	0.8047
BC	EfficientNet-B0, PNASNet, EfficientNet-B6	0.7989
BCD	EfficientNet-B2, EfficientNet-B5, ResNeSt101	**0.8132**
BD	EfficientNet-B0, EfficientNet-B5, PNASNet5Large	0.8100
CD	EfficientNet-B3, EfficientNet-B5, VGG19	0.8106

Across various ensemble experiments, with and without data augmentation, The "Average of 3" ensemble consistently outperforms both the "Average of all" and "Voting" ensembles in terms of balanced accuracy. It achieves higher balanced accuracy scores and consistently selects the best combination of 3 models from all possible model combinations, demonstrating its ability to leverage the strengths of these selected model combinations while excluding underperforming models.

It is worth noting that, for the "Average of 3" ensemble experiments with data augmentation, the models selected for the final combination are not necessarily the ones that best perform individually for each experiment. For example, for the BCD configuration, the models selected for the final ensemble were EfficientNet-B2, EfficientNet-B5, and ResNeSt101, but the models that better performed individually for this experiment were EfficientNet-B5, EfficientNet-B6, and PNASNet.

Figure 5.3 presents the best ensemble results for each experiment performed. By analyzing those results, we notice that the application of data augmentation and image preprocessing techniques improved the performance of the ensemble of models compared to the ensemble of the models trained with the raw data.

These results indicate that combining the predictions of multiple models can be beneficial by providing diverse perspectives and combining the strengths of each individual

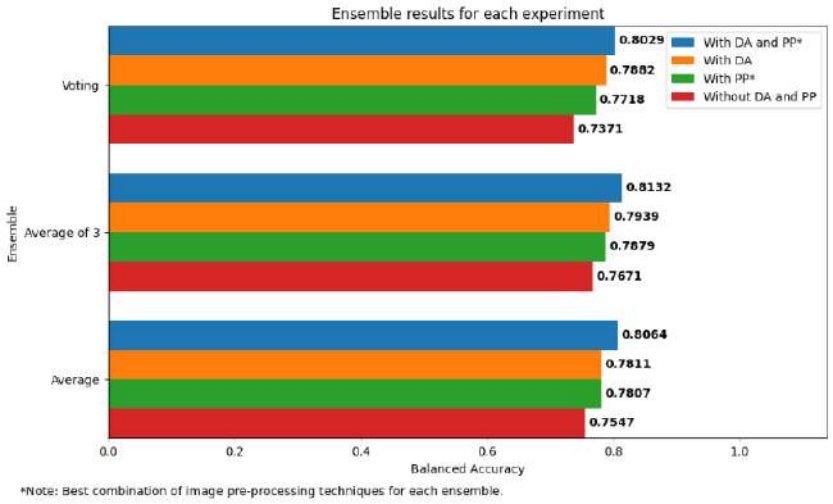


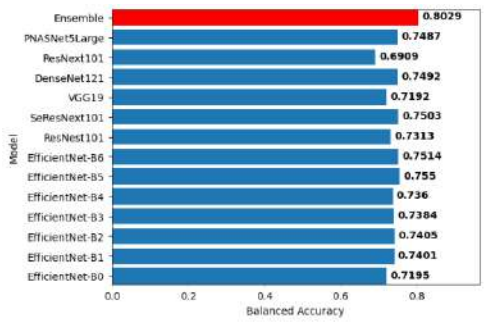
Figure 5.3: Ensemble results for each experiment.

model. Furthermore, the ensemble technique takes advantage of the diversity of the models and can reduce the impact of individual model weaknesses or errors.

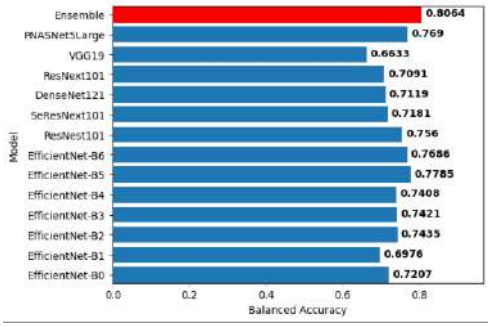
Table 5.7 presents the details of the best result for each ensemble technique and, in Figure 5.4, we can notice that the ensemble of the models performs better than the models individually.

Table 5.7: Configuration of best ensemble results

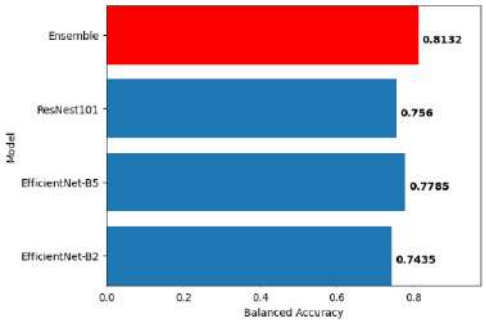
Ensemble	**Configuration**	**Models**	**Balanced accuracy**
Average of 3	Data augmentation, Color Constancy, Hair Removal and Center Crop	ResNeSt101, EfficientNet-B2, and EfficientNet-B5	0.8132
Average	Data augmentation, Color Constancy, Hair Removal and Center Crop	All models	0.8064
Voting	Data augmentation, Contrast Enhancement and Center crop	All models	0.8029



(a) Ensemble by voting.



(b) Ensemble average of all models.



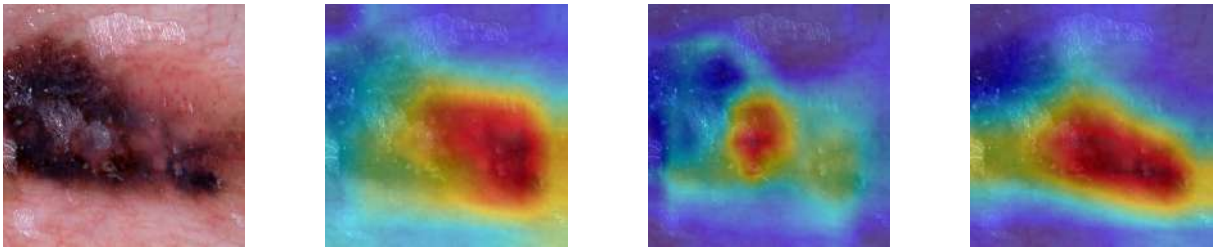
(c) Ensemble average of the best combination of 3 models.

Figure 5.4: Comparison of the results between the ensembles and each individual model.

Overall, the best-performing ensemble method was the "Average of 3" with the employment of data augmentation techniques and Color Constancy, Hair Removal, and Center Crop as image preprocessing steps, and resulted in a balanced accuracy of 0.8132. This model is the average of the results of the EfficientNet-B2, EfficientNet-B5, and ResNeSt101.

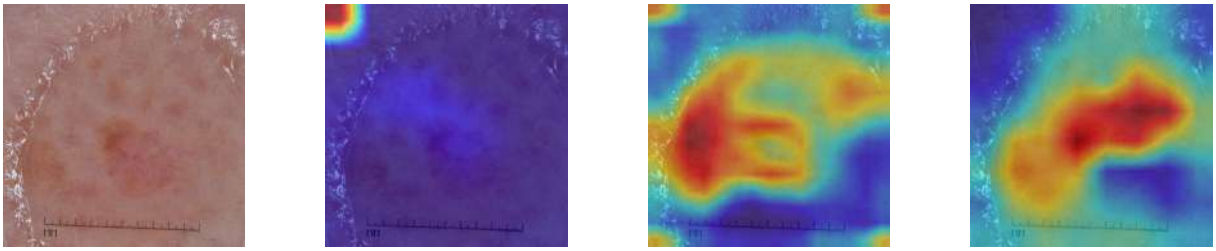
5.2 GradCAM insights

To gain insights into how these models are making their predictions, we employed the Gradient-weighted Class Activation Mapping (GradCAM) technique. This allowed us to visualize which parts of the input image each model is using to make its predictions, both in cases where the prediction is correct and in cases where it is wrong.



(a) Original image. (b) EfficientNet-B2 image. (c) EfficientNet-B5 image. (d) ResNeSt101 image.

Figure 5.5: GradCAM image visualization for a correct prediction of the images' class.



(a) Original image. (b) EfficientNet-B2 image. (c) EfficientNet-B5 image. (d) ResNeSt101 image.

Figure 5.6: GradCAM image visualization for a wrong prediction of the images' class.

In Figure 5.5, we present the visualization for cases where the models correctly classified the lesion, and we can observe that the decision was based on the part of the image that represents the lesion.

Figure 5.6 shows the visualization of cases where the models misclassified the lesion. It is evident that the models could not accurately identify the location of the lesion, and instead used unrelated information to classify it. It is worth noting that the lesion in the image is challenging to identify since the colors of the skin and lesion are very similar.

5.3 Comparison with other works

We compared our results to the three best models of the ISIC 2017 Challenge mentioned in Section 3, and we also assessed two other recent works that used the ISIC 2017 dataset. One approach introduced by Liu et al. [80] utilizes a Multi-level Relationship

Capture Network to comprehend relationships within and across images. Additionally, it incorporates a lesion discerning module and a consistency regularization module for feature extraction. The other approach proposed by Xie et al. [81] uses a mutual bootstrapping deep convolutional neural networks model for simultaneous skin lesion segmentation and classification and employs additional data from the PH2 dataset.

Table 5.8 presents the results of each approach mentioned. We evaluated the Area Under the ROC Curve (AUC), accuracy (ACC), sensitivity (SE), and specificity (SP) for melanoma and seborrheic keratosis classification. "Average AUC" denotes the mean value of AUC scores for both melanoma and seborrheic keratosis. The final metrics were obtained from the ISIC 2017 Leaderboard webpage.

Table 5.8: Comparison with other methods.

		Melanoma				Seborrheic Keratosis				
Method	**Average AUC**	**AUC**	**ACC**	**SE**	**SP**	**AUC**	**ACC**	**SE**	**SP**	**Balanced Accuracy**
[60] First place	0.9108	0.8681	0.8283	0.7350	0.8509	0.9535	0.8033	**0.9778**	0.7725	0.8311
[61] Second place	0.9101	0.8556	0.8233	0.1026	**0.9979**	0.9647	0.8750	0.1778	**0.9980**	**0.8833**
[62] Third place	0.9080	0.8735	0.8716	0.5470	0.9503	0.9425	0.8950	0.3556	0.9902	0.8440
[80] Liu et al.	**0.9680**	**0.9470**	**0.9060**	**0.8020**	0.9360	**0.9890**	**0.9490**	0.9180	0.9460	-
[81] Xie et al.	0.9380	0.9030	0.8780	0.7270	0.9150	0.9730	0.9300	0.8440	0.9450	-
Our method	0.9300	0.8957	0.8750	0.7094	0.9151	0.9643	0.9283	0.8778	0.9373	0.8132

Our proposed solution outperformed the top three solutions from the ISIC 2017 challenge in terms of Average AUC (0.9300) and melanoma AUC (0.8957). Additionally, it produced similar results to Xie et al. [81] without the need for segmentation or additional data. However, the method proposed by Liu et al. [80] outperformed our approach in almost every metric for both melanoma and seborrheic keratosis classification. This suggests that using more complex mechanisms for feature extraction can be a valuable direction for future research.

Chapter 6

Conclusions

In conclusion, our study on skin lesion classification has provided valuable insights into developing effective models for dermatological diagnosis. Throughout our research, we systematically examined various aspects of model performance, including individual models, image preprocessing, data augmentation, and ensemble techniques.

The image preprocessing techniques were particularly effective in standardizing the images and improving the quality of images. Our findings indicate that the use of image preprocessing techniques, such as Color Constancy, Hair Removal, and Center Crop, can enhance model performance. Notably, we observed substantial improvements with the application of these techniques, with certain models demonstrating sensitivity to specific preprocessing methods. However, they may also remove or modify some important features. Because of that, it is important to test several combinations for different models to obtain the best possible results.

Moreover, incorporating data augmentation proved to be a beneficial strategy. This approach significantly improved overall model performance. This technique can help to prevent overfitting and improve the models' robustness and ability to handle different image variations. However, the effectiveness of data augmentation largely depends on the types of transformations used. Some transformations may not be appropriate for skin lesion images and could even introduce misleading features.

The ensemble strategy effectively improved the overall performance by combining the

strengths of different models. However, the choice of models and the way their predictions are combined can significantly impact the performance of the ensemble.

Within our ensemble experiments, the "Average of 3" method consistently outperformed other ensemble strategies. This underscores its efficacy in selecting the most favorable combination of models for improved classification results, achieving a balanced accuracy of 0.8132.

Our model demonstrated great performance in the ISIC 2017 dataset. However, this may not translate to high performance in real-world scenarios since the used dataset may not capture all the variations and complexities of skin lesions encountered in clinical practice. Therefore, validating our models on additional datasets and for other skin lesion classes would be beneficial.

In our comparative analysis with the top-performing methods from the ISIC 2017 challenge and recent approaches, our method demonstrated competitive performance. It outperformed the top three ISIC 2017 solutions in Average AUC and melanoma AUC and achieved results similar to recent approaches, all without the need for intricate lesion segmentation or additional data. While Liu et al.'s method exhibited superior performance in certain metrics, our study suggests that there is potential for further advancements through the exploration of advanced feature extraction mechanisms.

In summary, our work represents significant insights into the task of skin lesion classification, emphasizing the importance of image preprocessing, data augmentation, and ensemble techniques. While our final solution may not significantly outperform all existing approaches, it contributes valuable insights and lays the foundation for continued research to enhance dermatological diagnosis and patient care.

Future investigations may explore advanced feature extraction techniques and innovative model architectures to elevate the capabilities of our system further and improve the accuracy of skin lesion classification. Additionally, the application of different image preprocessing techniques and more complex ensemble methods can be explored.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, pp. 209–249, 3 2021.

[2] J. Ferlay, M. Colombet, I. Soerjomataram, *et al.*, "Cancer statistics for the year 2020: An overview," en, *Int. J. Cancer*, vol. 149, no. 4, pp. 778–789, 2021.

[3] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA Cancer J. Clin.*, vol. 73, no. 1, pp. 17–48, 2023.

[4] D. Lazovich, K. Choi, and R. I. Vogel, "Time to get serious about skin cancer prevention," *Cancer Epidemiol. Biomarkers Prev.*, vol. 21, no. 11, pp. 1893–1901, 2012.

[5] T. L. Diepgen and V. Mahler, "The epidemiology of skin cancer," *Br. J. Dermatol.*, vol. 146, no. s61, pp. 1–6, 2002.

[6] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *Lancet Oncol.*, vol. 3, no. 3, pp. 159–165, 2002.

[7] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: Transforming the practice of medicine," *Future Healthcare Journal*, vol. 8, no. 2, e188–e194, 2021.

[8] N. C. Codella, D. Gutman, M. E. Celebi, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018*

IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, 2018, pp. 168–172.

[9] G. M. Silva, A. E. Lazzaretti, and F. C. Monteiro, "Deep learning techniques applied to skin lesion classification: A review," in *2022 International Conference on Machine Learning, Control, and Robotics (MLCR)*, 2022, pp. 106–111.

[10] G. M. Silva, A. E. Lazzaretti, and F. C. Monteiro, "An evaluation of image preprocessing in skin lesions detection," *III International Conference on Optimization, Learning Algorithms and Applications (OL2A 2023), Springer*, 2023.

[11] G. M. Silva, A. E. Lazzaretti, and F. C. Monteiro, "Deep learning techniques applied to skin lesion classification," *The 2023 2nd International Conference on Machine Learning, Control, and Robotics (MLCR 2023)*, 2023.

[12] T. Skuhala, V. Trkulja, M. Rimac, A. Dragobratović, and B. Desnica, "Analysis of types of skin lesions and diseases in everyday infectious disease practice—how experienced are we?" *Life*, vol. 12, no. 7, p. 978, 2022.

[13] W. Damsky and M. Bosenberg, "Melanocytic nevi and melanoma: Unraveling a complex relationship," *Oncogene*, vol. 36, no. 42, pp. 5771–5792, 2017.

[14] C. Hafner and T. Vogt, "Seborrheic keratosis," *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, vol. 6, no. 8, pp. 664–677, 2008.

[15] N. H. Matthews, W.-Q. Li, A. A. Qureshi, M. A. Weinstock, and E. Cho, "Epidemiology of melanoma," *Exon Publications*, pp. 3–22, 2017.

[16] M. Arnold, D. Singh, M. Laversanne, *et al.*, "Global burden of cutaneous melanoma in 2020 and projections to 2040," *JAMA dermatology*, vol. 158, no. 5, pp. 495–503, 2022.

[17] N. R. Abbasi, H. M. Shaw, D. S. Rigel, *et al.*, "Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria," *JAMA*, vol. 292, no. 22, pp. 2771–2776, 2004.

[18] G. Argenziano, C. Catricalà, M. Ardigo, *et al.*, "Seven-point checklist of dermoscopy revisited," *British Journal of Dermatology*, vol. 164, no. 4, pp. 785–790, 2011.

[19] L. R. Soenksen, T. Kassis, S. T. Conover, *et al.*, "Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images," *Sci. Transl. Med.*, vol. 13, no. 581, eabb3652, 2021.

[20] S. W. Menzies, "A method for the diagnosis of primary cutaneous melanoma using surface microscopy," *Dermatologic clinics*, vol. 19, no. 2, pp. 299–305, 2001.

[21] C. Rosendahl, A. Cameron, I. McColl, and D. Wilkinson, "Dermatoscopy in routine practice:'chaos and clues'," *Australian family physician*, vol. 41, no. 7, pp. 482–487, 2012.

[22] E. Errichetti and G. Stinco, "Dermoscopy in general dermatology: A practical overview," *Dermatology and therapy*, vol. 6, pp. 471–507, 2016.

[23] C. Carrera, M. A. Marchetti, S. W. Dusza, *et al.*, "Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma: A web-based international dermoscopy society study," *JAMA dermatology*, vol. 152, no. 7, pp. 798–806, 2016.

[24] S. Mane and S. Shinde, "A method for melanoma skin cancer detection using dermoscopy images," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India: IEEE, 2018.

[25] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, p. 420, 2021.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[28] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, p. 022 022, 2019.

[29] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[30] D. Mandic and J. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. Wiley, 2001.

[31] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[33] L. Alzubaidi, J. Zhang, A. J. Humaidi, *et al.*, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[34] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556 [cs.CV].

[35] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

[36] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2018. arXiv: 1608.06993 [cs.CV].

[37] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG].

[38] C. Liu, B. Zoph, M. Neumann, *et al.*, *Progressive neural architecture search*, 2018. arXiv: 1712.00559 [cs.CV].

[39] N. Codella, V. Rotemberg, P. Tschandl, *et al.*, *Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)*, 2019.

[40] M. S. K. C. Center, *The international skin imaging collaboration*, Last accessed 15 May 2022. [Online]. Available: https://www.isic-archive.com/.

[41] D. Gutman, N. C. F. Codella, E. Celebi, *et al.*, *Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the int. skin imaging collaboration (isic)*, 2016.

[42] P. Tschandl, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," 2018.

[43] M. Combalia, N. C. F. Codella, V. Rotemberg, *et al.*, *Bcn20000: Dermoscopic lesions in the wild*, 2019.

[44] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, p. 34, 2021.

[45] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "Ph2 - a dermoscopic image database for research and benchmarking," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2013.

[46] H. Liao, Y. Li, and J. Luo, "Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, IEEE, 2016.

[47] A. G. Pacheco, G. R. Lima, A. S. Salomão, *et al.*, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, p. 106 221, 2020.

[48] M. K. Hasan, M. A. Ahamad, C. H. Yap, and G. Yang, "A survey, review, and future trends of skin lesion segmentation and classification," *Computers in Biology and Medicine*, vol. 155, p. 106 624, 2023.

[49] A. Romero Lopez, X. Giro-i-Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *2017 13th IASTED International Conference on Biomedical Engineering (BioMed)*, 2017, pp. 49–54.

[50] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with alex-net," *PLoS One*, vol. 14, no. 5, e0217293, 2019.

[51] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imaging Graph.*, vol. 71, pp. 19–29, 2019.

[52] M. K. Hasan, M. T. E. Elahi, M. A. Alam, M. T. Jawad, and R. Marti, "DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation," *Inform. Med. Unlocked*, vol. 28, no. 100819, p. 100 819, 2022.

[53] M. A. Khan, M. Sharif, T. Akram, S. A. C. Bukhari, and R. S. Nayak, "Developed Newton-Raphson based deep features selection framework for skin lesion recognition," *Pattern Recognit. Lett.*, vol. 129, pp. 293–303, 2020.

[54] F. Afza, M. Sharif, M. Mittal, M. A. Khan, and D. Jude Hemanth, "A hierarchical three-step superpixels and deep learning framework for skin lesion classification," *Methods*, vol. 202, pp. 88–102, 2022.

[55] S. Khouloud, M. Ahlem, F. Touré, and A. Slim, "W-net and inception residual network for skin lesion segmentation and classification," *Applied Intelligence*, vol. 52, 2022.

[56] K. Jayapriya and I. J. Jacob, "Hybrid fully convolutional networks-based skin lesion segmentation and melanoma detection using deep feature," *International Journal of Imaging Systems and Technology*, vol. 30, no. 2, pp. 348–357, 2020.

[57] R. Sarkar, C. C. Chatterjee, and A. Hazra, "Diagnosis of melanoma from dermoscopic images using a deep depthwise separable residual convolutional network," *IET Image Process.*, vol. 13, no. 12, pp. 2130–2142, 2019.

[58] Y. Ge, B. Li, Y. Zhao, E. Guan, and W. Yan, "Melanoma segmentation and classification in clinical images using deep learning," in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, Macau China: ACM, 2018.

[59] A. Finnane, C. Curiel-Lewandrowski, G. Wimberley, *et al.*, "Proposed technical guidelines for the acquisition of clinical images of skin-related conditions," *JAMA dermatology*, vol. 153, no. 5, pp. 453–457, 2017.

[60] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, *Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble*, 2017. arXiv: 1703.03108 [cs.CV].

[61] I. G. Diaz, *Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions*, 2017. arXiv: 1703.01976 [cs.CV].

[62] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle, *Recod titans at isic challenge 2017*, 2017. arXiv: 1703.04819 [cs.CV].

[63] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, no. 100864, p. 100 864, 2020.

[64] A. G. C. Pacheco, G. R. Lima, A. S. Salomão, *et al.*, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," en, *Data Brief*, vol. 32, p. 106 221, 2020.

[65] Q. Ha, B. Liu, and F. Liu, "Identifying melanoma images using efficientnet ensemble: Winning solution to the simm-isic melanoma classification challenge," *arXiv preprint arXiv:2010.05351*, 2020.

[66] A. N. Hoshyar, A. Al-Jumaily, and A. N. Hoshyar, "The beneficial techniques in preprocessing step of skin cancer detection system comparing," *Procedia Comput. Sci.*, vol. 42, pp. 25–31, 2014.

[67] K. Lucknavalai and J. P. Schulze, "Real-time contrast enhancement for 3d medical images using histogram equalization," in *Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part I 15*, Springer, 2020, pp. 224–235.

[68] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, "DullRazor: A software approach to hair removal from images," *Comput. Biol. Med.*, vol. 27, no. 6, pp. 533–543, 1997.

[69] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, 2019.

[70] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, 2016.

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

[72] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[73] C. Liu, B. Zoph, M. Neumann, *et al.*, "Progressive neural architecture search," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

[74] H. Zhang, C. Wu, Z. Zhang, *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2736–2746.

[75] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016.

[76] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[77] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[78] F. Perez, S. Avila, and E. Valle, "Solo or ensemble? choosing a cnn architecture for melanoma classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[79] A. Pacheco, *Raug*, `https://github.com/paaatcha/raug`, 2020.

[80] Z. Liu, R. Xiong, and T. Jiang, "Multi-level relationship capture network for automated skin lesion recognition," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, *et al.*, Eds., Springer International Publishing, 2021, pp. 153–164.

[81] Y. Xie, J. Zhang, Y. Xia, and C. Shen, *A mutual bootstrapping model for automated skin lesion segmentation and classification*, 2020. arXiv: `1903.03313 [cs.CV]`.