Deep Learning - Project Report

Group 22

Enhancing Diagnostic Accuracy with AI
Advanced Image Analysis for Skin Disease Identification

Authors:

Guilherme Sá	20230520
Helena Mashayekhi	20230561
Raquel Rocha	20230188
Sebastião Rosalino	20230372
Zenan Chen	20221390

NOVA IMS

Academic Orientation: Mauro Castelli and Yuri Perezhohin

2023/2024

Contents

Introduction

In the realm of medical diagnostics, precision and efficiency are crucial, particularly in dermatology where visual assessment is fundamental. This project harnesses the power of Deep Learning to enhance the Accuracy of skin disease classification, a critical advancement given the subtle distinctions between various skin conditions. Utilizing advanced convolutional neural networks, specifically pre-trained models like DenseNet, VGG, and ResNet, this study aims to automate the analysis of dermatological images, offering a robust tool to aid healthcare professionals.

The integration of Gradient-weighted Class Activation Mapping (Grad-CAM) provides further insights by highlighting key areas in images that influence the model's predictions, thus adding a layer of interpretability to the AI's decision-making process. Through model training and validation, coupled with extensive hyperparameter optimization using Optuna, this project strives to bridge the gap between clinical dermatology and artificial intelligence, potentially transforming diagnostic procedures by providing swift, accurate, and interpretable results.

Data Exploration and Preprocessing

The dataset of images was taken from the "FITZPATRICK17" repository. It comprises several features: an image URL, nine-partition and three-partition labels, two human skin color scales (Fitzpatrick Scale and Fitzpatrick Centaur Scale), a quality control feature (qc), and the disease label, which serves as the target variable. The qc feature provides doctor assessments on approximately 3% of the records, indicating the correctness of some classifications. The final dataset includes 16,577 records across ten columns as displayed in Appendix A, Table 1.

Missing values were found in the qc (16,073 missing) and URL (41 missing) features. Records lacking a URL were removed as they contained no retrievable image data. Descriptive analysis of the predictors (Table 2) led to the exclusion of records flagged as 'Wrongly labeled' in the qc feature to prevent training on incorrectly classified data. The distribution of the skin color scales and the three-partition label revealed a predominance of non-neoplastic cases (72.9%), followed by malignant (13.7%) and benign cases (13.5%), as shown in Appendix B, Figure 2. Detailed distributions of the nine-partition label feature correlated with their respective three-partition labels are shown in Appendix B, Figure 3. Notably, the most prevalent condition is psoriasis (653 cases), while xanthomas, pustular psoriasis, and pilomatricoma are the least common, each with 53 cases. Detailed information regarding the distribution of the diseases can be visualized in Figure 4 and Figure 5.

Irrelevant features such as 'md5hash', 'qc', 'url_alphanum', and 'three_partition_label' were removed (the latter turned out to be redundant since all its information was already being provided by nine-partition label). Observations where both Fitzpatrick Scales were marked as -1 (indicating no measurement) were also discarded. A new feature, 'fitzpatrick_average', was calculated as the average of the two scales, except when one scale read -1, in which case the other scale's value was used exclusively. The distributions of the original Fitzpatrick scales are shown in Appendix B, Figure 1.

The images were then converted into 128x128 RGB matrices (the quality had to be lowered due to

RAM constraints). This transformation was executed before splitting the dataset into 60% for training, 20% for validation, and 20% for testing. Each subset was stratified to reflect the original distribution of labels, ensuring that each set was representative of the overall dataset.

To prevent data leakage, normalization of the 'fitzpatrick_average' feature was carried out using a MinMaxScaler fitted exclusively on the training data, as its distribution was non-normal (Appendix C, Figure 6). This scaler adjusted values to a 0-1 range.

The 'nine_partition_label' was one-hot encoded to facilitate its inclusion in model predictions that use a Functional API.

Data augmentation techniques were employed to address class imbalances in the training dataset, specifically horizontal and vertical flipping, along with zooming up to 20%, as can be observed in Appendix D, Figure 7 and Figure 8. This approach, recommended by Buda, M. et al. (2018), ensured equal representation across classes. Each class was augmented to initially comprise 3% of the training set. However, as the dataset expanded, each class ended up representing about 0.9% of the total. Moreover, the 'fitzpatrick_average' and nine-partition label features were replicated for augmented records, as these attributes do not change with image augmentation. Post-augmentation, the dataset was randomly shuffled to promote better model convergence and to prevent the models from learning any order-specific patterns.

To enhance model performance for specific diseases, the project explored a method to digitally remove body hair from images. This technique aimed to create cleaner images by reducing visual noise caused by hair, simplifying disease identification. Examples of these processed images are displayed in Appendix D, as shown in Figure 9 and Figure 10, with a detailed description of the method available in the bibliography (sunnyshah289, 2018). However, the results from images without body hair did not meet expectations, leading to the discontinuation of this approach. The results are displayed on Appendix E, Figure 3, Figure 4 and Figure 5. It is believed that the quality of the images was adversely affected by the hair removal process, often degrading the clarity of other critical features within the images.

Experimental Setup

This section details the experimental setup used for the development of the models, emphasizing the integration of advanced machine learning techniques and tools.

Pretrained Models: This project leveraged pre-trained models such as VGG16, DenseNet201, and ResNet50. These models were initially trained on the ImageNet dataset, enabling the utilization of their learned features for the specific task of dermatological image classification. For all pre-trained models, their top layer was not included and was replaced by a Dense layer of 114 units, since the current task is a 114 multi-class problem. Furthermore, the input shape was altered to 128x128x3 as this is the shape of the pictures. Depending on the pre-trained model, it was decided to unfreeze a different number of top layers to fine-tune them.

Optuna for Hyperparameter Tuning: Optuna, an optimization framework, was used to systematically explore and identify the best hyperparameters for the models. Optuna's efficient search capability allowed for the fine-tuning of the learning rates, optimizers, layer configurations and batch sizes.

Top-K Accuracy Metric: It was decided to, upon the Accuracy and Weighted F1-Score, employ the Top

K Accuracy Score metric provided by Scikit-Learn Metrics sub-module. This metric evaluates whether the true label is within the top 'K' predictions provided by the model, addressing the complexity of the multi-class classification task by offering a narrower evaluation scope. This approach not only provides a more nuanced assessment of model accuracy but also supports clinical decision-making by presenting multiple probable diagnoses, thereby aiding healthcare professionals in narrowing down diagnostic options.

Model Training Callbacks: A set of Keras' API callbacks was consistently utilized to prevent overfitting. These included the Model Checkpoint which saved the model after each epoch if it showed improvement in terms of validation loss, ensuring that the best-performing model was always saved. The Early Stopping callback was employed to halt training when the validation loss ceased to improve for 5 epochs, saving computational resources and preventing overfitting. Additionally, the Reduce Learning Rate On Plateau callback automatically reduced the learning rate when the validation loss plateaued for 3 epochs, providing a strategy to fine-tune the models under plateau conditions without manual intervention.

Approaches

To optimize performance in dermatological image classification, a series of eight models were evaluated, ranging from simple architectures to sophisticated pre-trained models. Performance metrics included Accuracy, Weighted F1-Score, Top 3 accuracy, Top 10 accuracy, detailed in Appendix E, Table 4. Each model's architecture concluded with a softmax-activated dense layer featuring 114 units for disease categorization and the loss function was the Categorical Cross Entropy.

The initial model, a basic configuration with a single convolutional layer, achieved minimal success, recording a Test Accuracy of only 7.6% and a Top 3 Accuracy of 16.8%. Subsequent enhancements included additional convolutional layers with increasing filters, batch normalization, and max pooling, which improved pattern recognition capabilities. Dropout regularization was also incorporated, resulting in the 'Regularized Model' that improved Test Accuracy to 15.8% and Top 3 Accuracy to 27.5%.

Advanced models utilized Keras' Functional API to accommodate multi-input formats, integrating 'fitzpatrick_average' and one-hot-encoded nine-partition labels. This approach yielded a significant performance increase, with the 'Functional API with Data Augmentation' model achieving a Test Accuracy of 32.2% and a Top 3 Accuracy of 52.6%. Pre-trained models were specifically adapted by replacing their top layers, originally designed for ImageNet classifications, with custom layers suited to the 114-class dermatology dataset. VGG16, initiated without data augmentation and without Functional API, displayed limited effectiveness, leading to its exclusion from further augmentation trials.

In contrast, DenseNet201 demonstrated substantial promise; particularly, the version incorporating data augmentation and Functional API significantly outperformed others, achieving the highest Test Accuracy of 47.7% and Top 3 Accuracy of 68.1%. Despite similar augmentation and API integration, ResNet50's performance lagged, suggesting a possible mismatch with the specific demands of dermatological imagery, as evidenced by its lower Test Accuracy of 25.7%.

DenseNet201's superior results underscore its effectiveness, attributed to an optimal unfreezing strategy (unfreezing layers past 'conv5_block30_0_bn') that allowed deeper network layers to adapt more effectively to dermatological features. This model, detailed further in the Results section, sets a bench-

mark for future enhancements in dermatological image classification.

Results

Using the pre-trained DenseNet201 with Functional API and data augmentation, the best performance achieved on the test set was a F1-Score of 0.47, an Accuracy of 0.48 and the top 3 K Accuracy of 0.68. The validation loss amounted to 2.12.

The best performing hyperparameters were: '`dropout_rate`': ≈ 0.44, '`num_units`': 64, '`num_units_one_hot`': 16, '`dropout_rate_combined`': ≈ 0.60, '`optimizer`': 'SGD', '`learning rate`': ≈ 0.03, '`batch_size`': 32.

As presented in Appendix F, Figure 11 (Error Analysis and Confusion Matrix), the labels acne vulgaris (label 2), basal cell carcinoma (8), folliculitis (32), hailey hailey disease (36), kaposi sarcoma (42), melanoma (56), mycosis fungoides (60), nematode infection (64), neurofibromatosis (66), neutrophilic dermatoses (68), pityriasis rubra pilaris (80), porokeratosis actinic (81), psoriasis (86), scabies (92) and squamous cell carcinoma (98) stand out for their amount of correct classifications.

However, it can be pointed out that for the diseases: pilomatricoma (77) and striae (101) no right classification was found, meaning that the model's ability to capture those diseases' patterns is lacking.

Furthermore, when looking at the remaining Confusion Matrix (Figure 11), the most significant errors were observed in the labels basal cell carcinoma (8), psoriasis (86), and squamous cell carcinoma (98). Potential reasons for these errors are body hair that distract from the disease, suboptimal image resolution that obscures critical details and the presence of blood that could distract from key diagnostic features. Examples for misclassified diseases can be observed in Figure 12. Misclassifications of diseases such as basal cell carcinoma (8) and squamous cell carcinoma (98) are particularly concerning because they are categorized as 'malignant' in three-partition label, highlighting the severe implications of such errors. In contrast, the misclassification of psoriasis (86), which falls under the 'non-neoplastic' category in the three-partition label, is considered less detrimental.

The classification performance for basal cell carcinoma (8) and squamous cell carcinoma (98) on the test dataset reveals a pattern of both high correct classification rates and significant misclassification rates. Specifically, the confusion matrix records 53 correct classifications and 38 misclassifications for basal cell carcinoma (8), and 72 correct identifications versus 36 misclassifications for squamous cell carcinoma (98). This pattern indicates that the model has successfully learned certain features of these two carcinoma types, yet it still confuses deciding between the two referred types. This issue may arise from an insufficient variety of examples in the training data that fully capture the diversity of each carcinoma type.

Visualizing Decisions in Dermatological Diagnostics with Grad-CAM

The best model, built on a Functional API with multiple types of inputs, was incompatible with Gradient-weighted Class Activation Mapping. Therefore, the analysis was conducted using the best performing single-input model, specifically a DenseNet model without Functional API and without data augmentation. For this purpose, the model's last convolutional layer was used because it aggregates the most

comprehensive information from prior layers, capturing high-level features responsible for making predictions. Only correct classifications were plotted to ensure relevance on the visual explanations provided.

Grad-CAM provides a visual explanation for the decisions made by convolutional networks. It does so by creating heatmaps superimposed on the original images, highlighting areas critical to the network's predictions. In these heatmaps, warm colors such as reds and oranges indicate regions with high influence on the output decision, while cool colors like blue denote areas of lesser importance. This method allows for a side-by-side visualization of the original images, the heatmaps, and the superimposed images as detailed in Appendix G, Figure 13.

Future Work

The promising results, characterized by almost 50% Accuracy and 85% of Top 10 Accuracy, as well as nuanced diagnostic suggestions, demonstrate the model's potential as a valuable tool for aiding in dermatological diagnosis, establishing a robust baseline for future work. The focus areas of improvement should be the following:

More Optuna trials with data augmentation: Given additional time, an expansion of Optuna trials for pre-trained models using data augmentation and Functional API techniques would be conducted. Results using these methodologies were promising, suggesting that a more thorough exploration of hyperparameters might uncover even superior configurations. This expanded search would aim to optimize the models further, potentially enhancing their generalizability.

Different unfreezing strategies for pre-trained models: Another area of potential improvement involves experimenting with various strategies for unfreezing layers in pre-trained models. Adjusting the depth of layer unfreezing could significantly affect the ability of the models to learn more complex patterns and generalize better on unseen data. More time would allow for systematic testing of these strategies to identify the most effective approach for each model type.

More Advanced data augmentation techniques: Exploration of additional data augmentation techniques, such as adjustments in brightness and contrast, could be beneficial. These methods could help in creating a more robust model by training on images that closer mimic the variability found in real-world scenarios. Enhancing exposure to different lighting conditions and contrasts could improve diagnostic correctness.

Requisition of dataset balance regarding skin response to UV light: The dataset featured an average Fitzpatrick skin type of 2.65, on a scale from 1 to 6. A more balanced dataset in terms of skin response to ultraviolet light could potentially improve model performance, as the models would train on a representative spectrum of skin types. Engaging with data providers to acquire or generate a dataset with a more even distribution across the Fitzpatrick scales could enhance the model's utility across different demographic groups. This approach would not only improve model fairness but also its applicability in diverse clinical settings.

Bibliography

Buda, M., Maki, A., & Mazurowski, M. A. (2018).A systematic study of the class imbalance problem in convolutional neural networks. Neural networks, 106, 249-259.

sunnyshah2894.(2018, August 14).DigitalHairRemoval. https://github.com/sunnyshah2894/Digital-HairRemoval/blob/master/README.md

Appendix

Appendix A - Dataset Features

Column name	Observations
md5hash	Identifier
fitzpatrick_scale	Contains AI generated classification of skin colour on a scale of 1 to 6
fitzpatrick_centaur	Contains a different AI generated classification of skin colour on a scale of 1 to 6
label	Contains the 114 dermatological disease classifications (i.e. psoriasis)
nine_partition_label	Contains a classification into 9 labels from the dermatological diseases' (i.e. inflammatory)
three_partition_label	Contains a classification into 3 labels from the dermatological diseases' (i.e. non-neoplastic)
qc	Contains the quality control
url	Contains the url that leads to the pictures
url_alphanum	Contains a version of the URL with non-alphanumeric characters removed
image_path	Contains the path to the stored images

Table 1: Feature types

Feature	Count	Unique	Top	Freq	Mean	Std	Min	25%	50%	75%	Max
md5hash	16577	16577	5e82a45bc5d7	1	-	-	-	-	-	-	-
fitzpatrick_scale	16577	-	-	-	2.69	1.53	-1	2	2	4	6
fitzpatrick_centaur	16577	-	-	-	2.26	1.66	-1	1	2	3	6
label	16577	114	psoriasis	653	-	-	-	-	-	-	-
nine_partition_label	16577	9	inflammatory	10886	-	-	-	-	-	-	-
three_partition_label	16577	3	non-neoplastic	12080	-	-	-	-	-	-	-
qc	504	5	1 Diagnostic	348	-	-	-	-	-	-	-
url	16536	16536	https://www.derm...	1	-	-	-	-	-	-	-
url_alphanum	16577	16577	httpwwwdermaami...	1	-	-	-	-	-	-	-
image_path	16525	16525	images/httpwww...	1	-	-	-	-	-	-	-

Table 2: Statistical Summary of Dataset Features

Appendix B - Data Exploration Plots

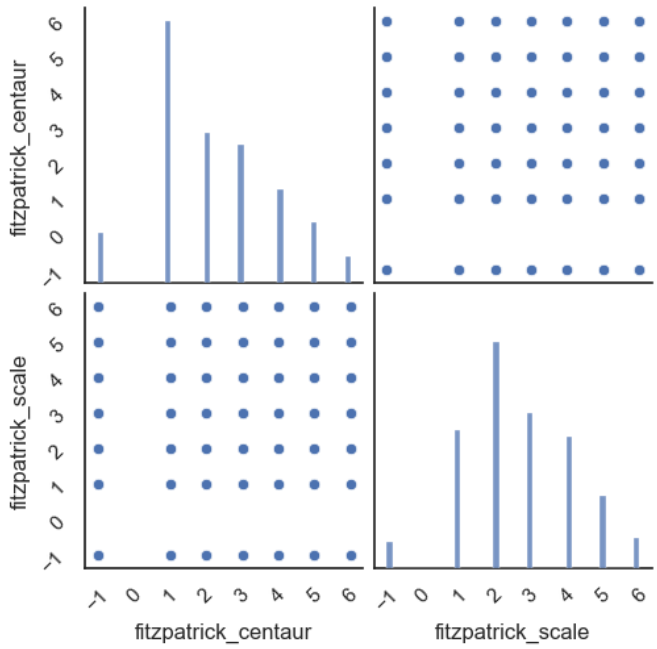

Figure 1: Fitzpatrick features distribution

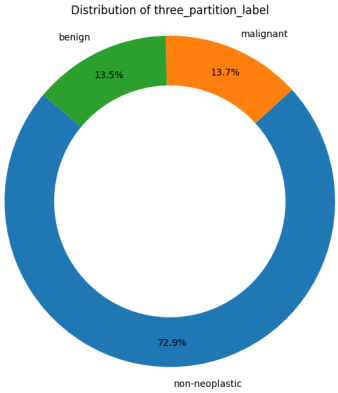

Figure 2: Distribution of Three-Partition Label

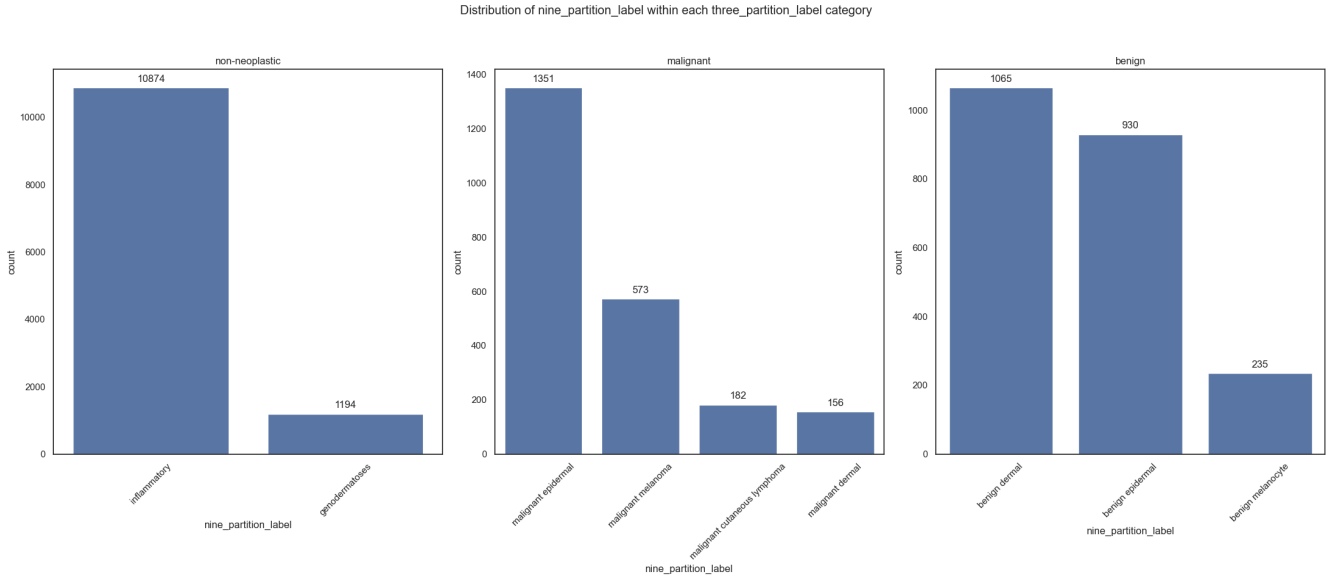

Figure 3: Distribution of Nine Partition Label within Three Partition Label

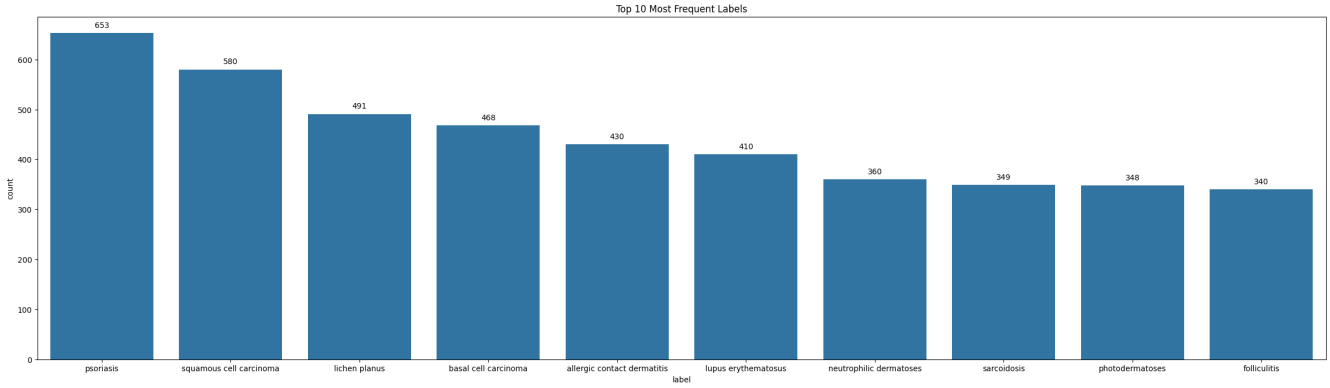

Figure 4: Top 10 Most Frequent Labels

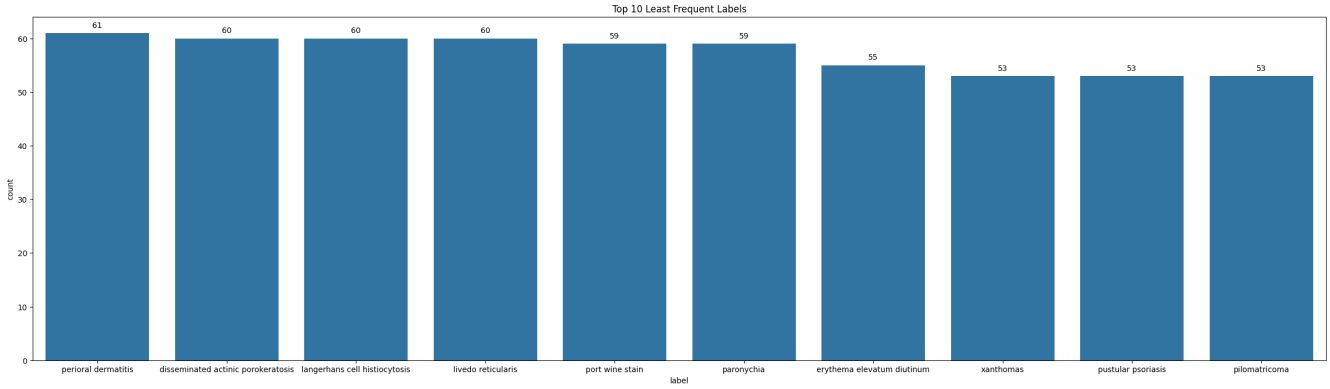

Figure 5: Top 10 Least Frequent Labels

Appendix C - Fitzpatrick Average Distribution

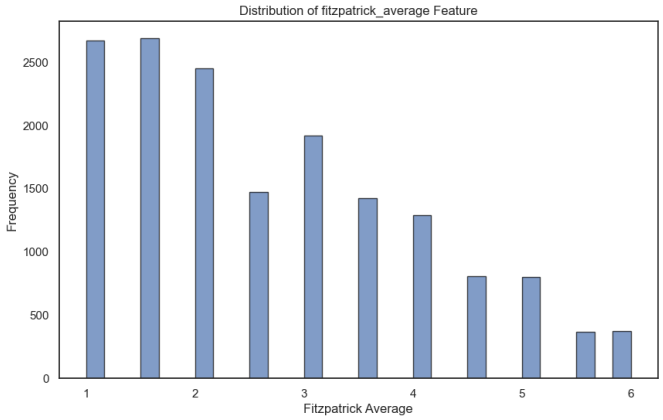

Figure 6: 'fitzpatrick_average' feature distribution

Appendix D - Data Augmentation Results and Hair Removal

Figure 7: Augmented images

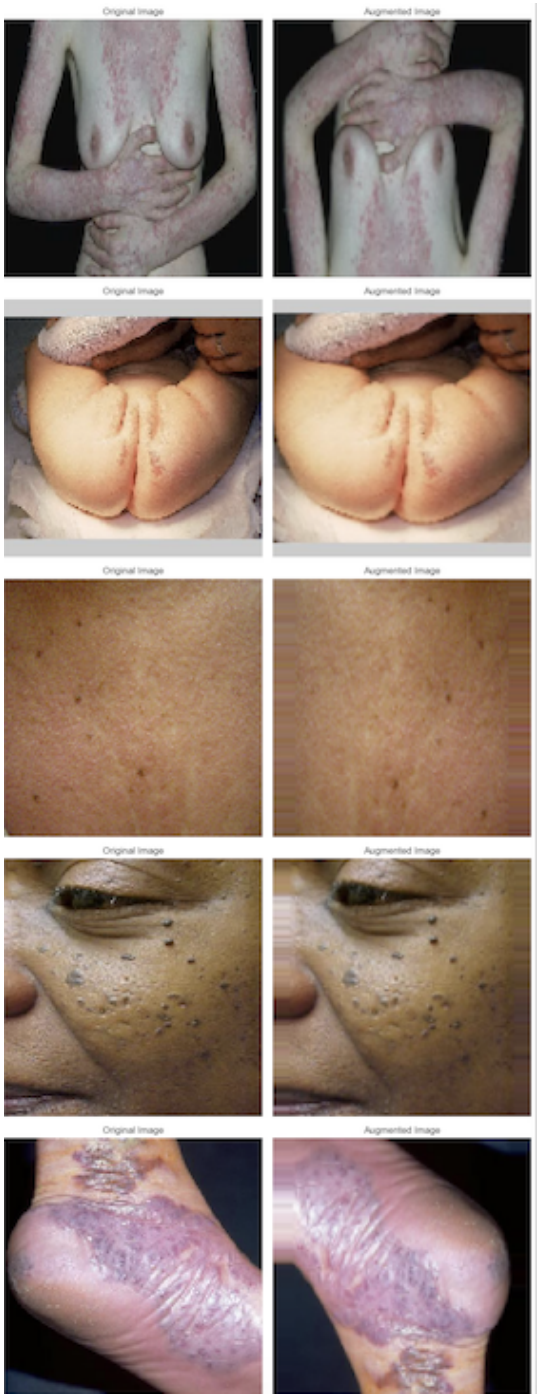

Figure 8: Augmented images

Figure 9: Hair Removal

Figure 10: Hair Removal

Appendix E - Models Parameters and Results

Model	Total params	Trainable params	Non-trainable params
Base Model	18,099,034 (69 MB)	18,099,034 (69 MB)	0 (0 Byte)
Regularized Model	12,998,194 (50 MB)	12,997,746 (50 MB)	448 (2 KB)
Functional API without Data Augmentation	13,067,058 (50 MB)	13,065,330 (50 MB)	1,728 (7 KB)
Functional API with Data Augmentation	13,067,058 (50 MB)	13,065,330 (50 MB)	1,728 (7 KB)
VGG 16 Without Functional API without Data Augmentation	15,648,690 (60 MB)	15,388,530 (59 MB)	260,160 (1,016 KB)
DenseNet 201 Without Functional API without Data Augmentation	18,540,978 (71 MB)	1,058,034 (4 MB)	17,482,944 (67 MB)
DenseNet201 With Functional API with Data Augmentation	19,415,314 (74 MB)	1,927,346 (7 MB)	17,487,968 (67 MB)
ResNet50 With Functional API with Data Augmentation	24,747,090 (94 MB)	6,674,482 (25 MB)	18,072,608 (69 MB)

Table 3: Number of parameters of each Model

Model	Test Accuracy	Test F1-Score	Top 3 Accuracy	Top 10 Accuracy	Validation Loss
Base Model	7.60%	5.30%	16.8%	36%	4.23009
Regularized Model	15.8%	14.3%	27.5%	47%	4.05701
Functional API without Data Augmentation	32.3%	31.1%	50.6%	73%	2.92900
Functional API with Data Augmentation	32.2%	31.6%	52.6%	75%	2.64600
VGG 16 Without Functional API without Data Augmentation	19.5%	17.7%	35.0%	57%	3.58679
DenseNet 201 Without Functional API without Data Augmentation	32.4%	31.8%	50.2%	70%	3.85702
DenseNet201 With Functional API with Data Augmentation	47.7%	47.1%	68.1%	85%	2.11500
ResNet50 With Functional API with Data Augmentation	25.7%	24.5%	45.8%	69%	2.87300

Table 4: Results for each Model with the Best Parameters

Model	Origal Images Test Accuracy	Hair Removed Images Test Accuracy	Origal Images Validation Loss	Hair Removed Images Validation Loss
Base Model	7.60%	10.6%	4.23009	4.141312
Regularized Model	15.8%	16.4%	4.05701	4.07221
Functional API with Data Augmentation	32.2%	31.5%	2.64600	2.82920
DenseNet201 With Functional API with Data Augmentation	47.7%	43.6%	2.11500	2.21550

Table 5: Comparison of Models Results with and without hair removal with best parameters

Appendix F - Confusion Matrix and Error Analysis

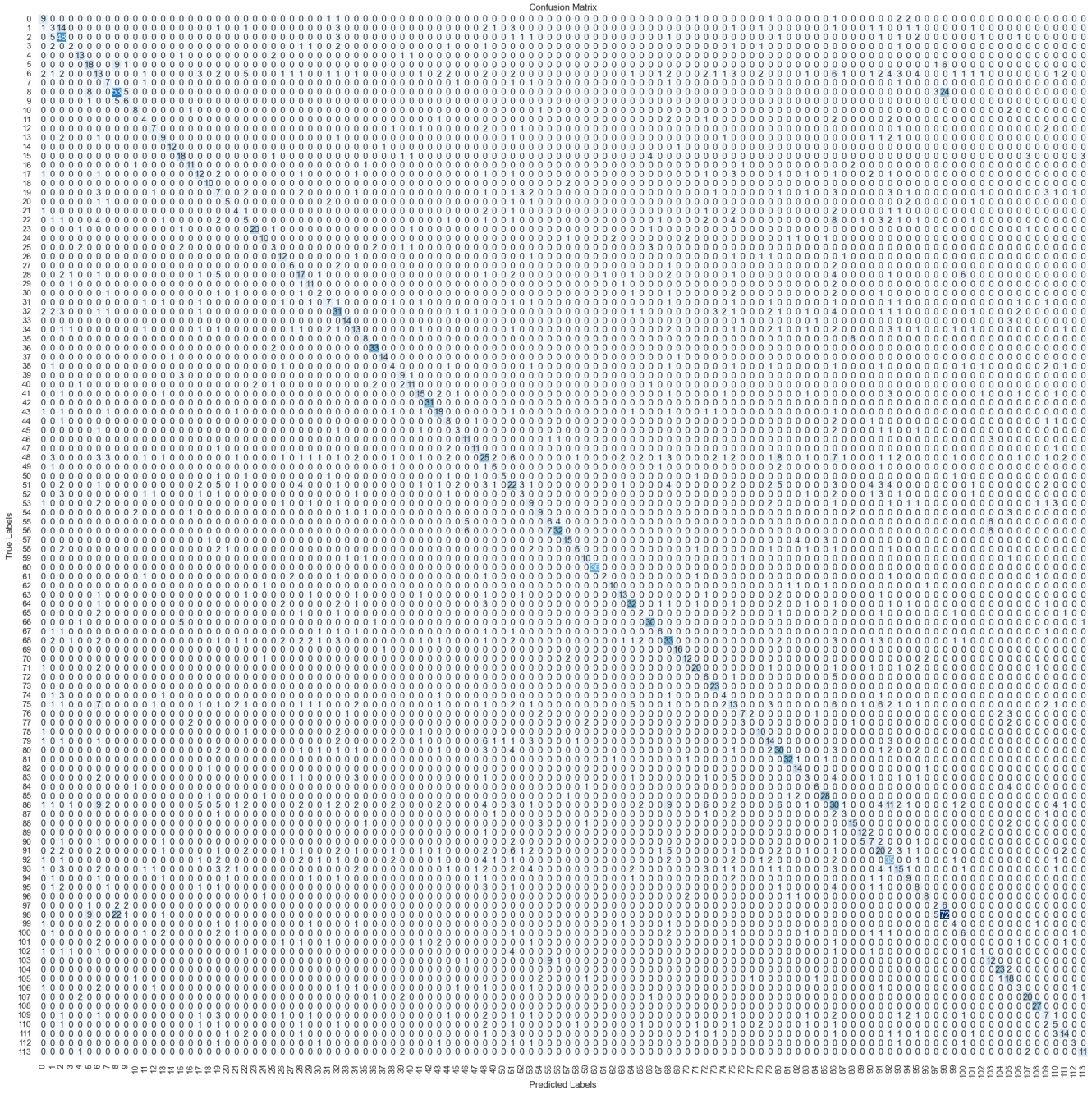

Figure 11: Confusion Matrix Visualization

Figure 12: Misclassified Images

Appendix G - Grad-CAM

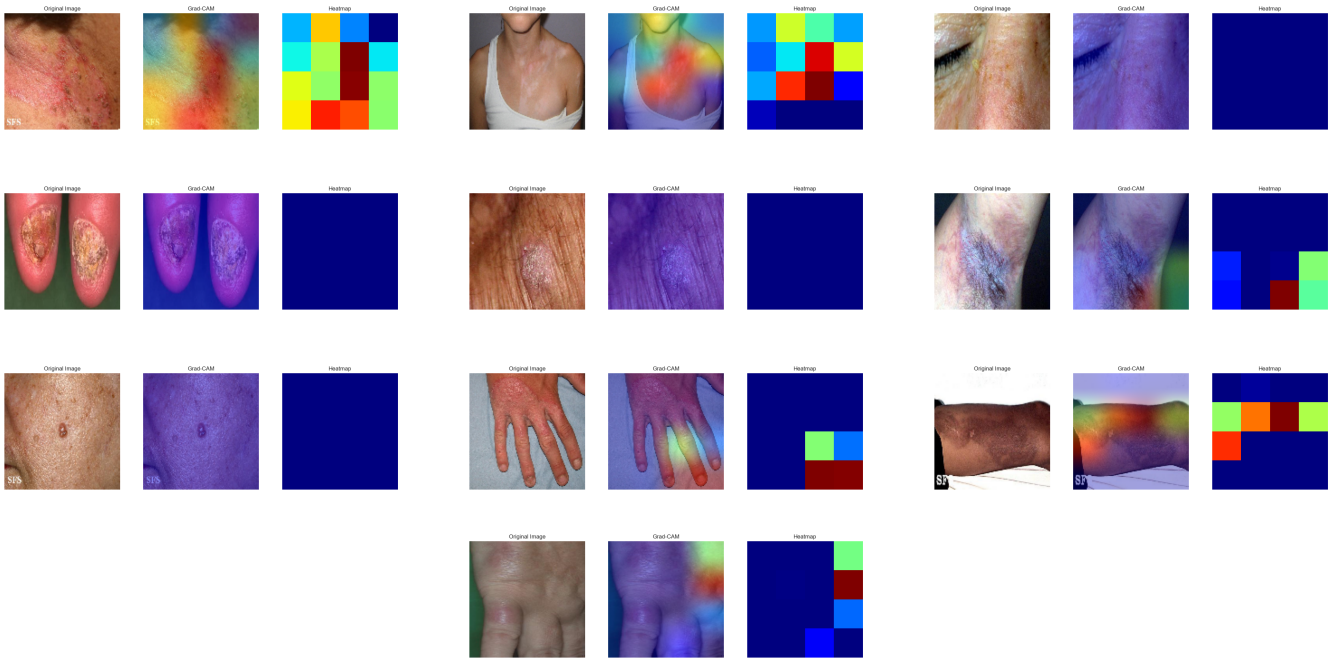

Figure 13: A collection of Grad-CAM images