

# Skin Lesions Classification: An Efficient MultiClass Skin Cancer Classification Using Transfer Learning

Ofir Neshher<sup>1</sup>, Yuval Katz<sup>1</sup>, Nadav Loebel<sup>2</sup>, Anna Aronovich<sup>3</sup>, Alon Oring<sup>1</sup>, and  
Zohar Yakhini<sup>1</sup>

<sup>1</sup>Department of Computer Science, Reichman University

<sup>2</sup>Head of AI, Beilinson Innovation

<sup>3</sup>Department of Dermatology, Rabin Medical Center (Beilinson Hospital)

June 25, 2023

The code for the project can be found in the GitHub repository available at

[https://github.com/Neshher123/skin\\_lesions\\_classification](https://github.com/Neshher123/skin_lesions_classification)

## Abstract

Skin cancer is the most common cancer in the world, constituting one-third of cancer cases. Benign skin cancers are not fatal and can be cured with proper medication. But it is not the same as malignant skin cancers. In the case of malignant melanoma, in its peak stage, the maximum life expectancy is less than or equal to 5 years. But, it can be cured if detected in the early stages. Though there are numerous clinical procedures, the accuracy of diagnosis falls between 49% to 81% and is time-consuming. So, dermatoscopy has been brought into the picture. It helped increase the accuracy of diagnosis but could not demolish the error-prone behavior. A quick and less error-prone solution is needed to diagnose this majorly growing skin cancer. This project deals with the usage of deep learning in skin lesion classification. In this project, an automated model for skin lesion classification using dermoscopic images is developed with CNN (Convolution Neural Networks) as a training model. Convolution neural networks are known for capturing the features of an image. So, they are preferred in analyzing medical images to find the characteristics that drive

the model toward success. This work has the potential to assist dermatology specialists in decision-making at critical stages.[1]

## 1 Introduction

Skin lesions can manifest in various forms and have different underlying causes. The ability to accurately diagnose a skin lesion can, sometimes, save lives as the prognosis is vastly depends on the time of diagnosis. The main categories of skin lesions are (The origin of the information is US National Library of Medicine National Institutes of Health):

**Melanocytic nevi** - Benign (non-cancerous) skin growths that are caused by an overgrowth of pigment-producing cells (melanocytes) in the skin. They are commonly referred to as moles and can range in color from brown to black, with a smooth or slightly raised surface. Most people have a few moles and they are typically harmless, but changes in size, shape, color, or texture can be an indication of skin cancer, so it is important to have them regularly checked by a doctor.

**Melanoma** - The most serious type of skin cancer, develops in the cells (melanocytes) that produce melanin - the pigment that gives your skin its color. Melanoma can also form in your eyes and, rarely, inside your body, such as in your nose or throat.

**Benign keratosis** - A type of noncancerous skin growth that appears as a scaly or wart-like bump on the surface of the skin. It is commonly found on sun-exposed areas such as the face, neck, arms, and legs. Benign keratosis is usually harmless, but in some cases, it may become irritated or infected. Treatment options include topical creams, cryotherapy, or surgical removal.

**Basal cell carcinoma** - A skin cancer that affects basal cells, which are the cells in the lower part of the outer layer of the skin. It is the most common form of skin cancer and often appears as a small, shiny bump or nodule, or as a flat, scaly area on the skin. BCC is typically slow-growing and does not usually spread to other parts of the body.

**Actinic keratosis** - A skin lesion that appears as a scaly or crusty bump on sun-exposed skin, such as the face, ears, neck, arms, and hands. It is caused by long-term exposure to ultraviolet (UV) radiation from the sun or indoor tanning beds. They are not cancerous, but a small

fraction of them will develop into skin cancer over time if left untreated. Treatment options include topical creams, cryotherapy (freezing), and surgical removal.

**Vascular lesions** - Skin conditions that are caused by abnormal blood vessels in the skin. They can be benign (non-cancerous) or malignant (cancerous).

**Dermatofibroma** - A benign skin growth that appears as a firm, raised bump on the skin. It is usually brown or reddish in color and can be located anywhere on the body, but is most commonly found on the legs. Dermatofibromas are usually harmless and do not require treatment, but they can be removed for cosmetic reasons if desired. Dermatofibromas are caused by an overgrowth of a mixture of different cell types in the dermis layer of the skin.

## 2 Method

### 2.1 Dataset

The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: More than 50% of lesions are confirmed through histopathology ('histo'), and the ground truth for the rest of the cases is either follow-up examination ('follow up'), expert consensus ('consensus'), or confirmation by in-vivo confocal microscopy ('confocal'). HAM10000 dataset is a benchmark dataset with over 50% of lesions confirmed by pathology.[2]

	Domain	Class	Total images	%
1	Melanocytic nevi	nv	6705	66.94%
2	Melanoma	mel	1113	11.11%
3	Benign keratosis	bkl	1099	10.97%
4	Basal cell carcinoma	bcc	514	5.13%
5	Actinic keratosis	akiec	327	3.26%
6	Vascular	vasc	142	1.41%
7	Dermatofibroma	df	115	1.14%
Total			10015	100%

Table 1: Dataset distribution



Figure 1: Sample images from the HAM10000 data-set for cancer types

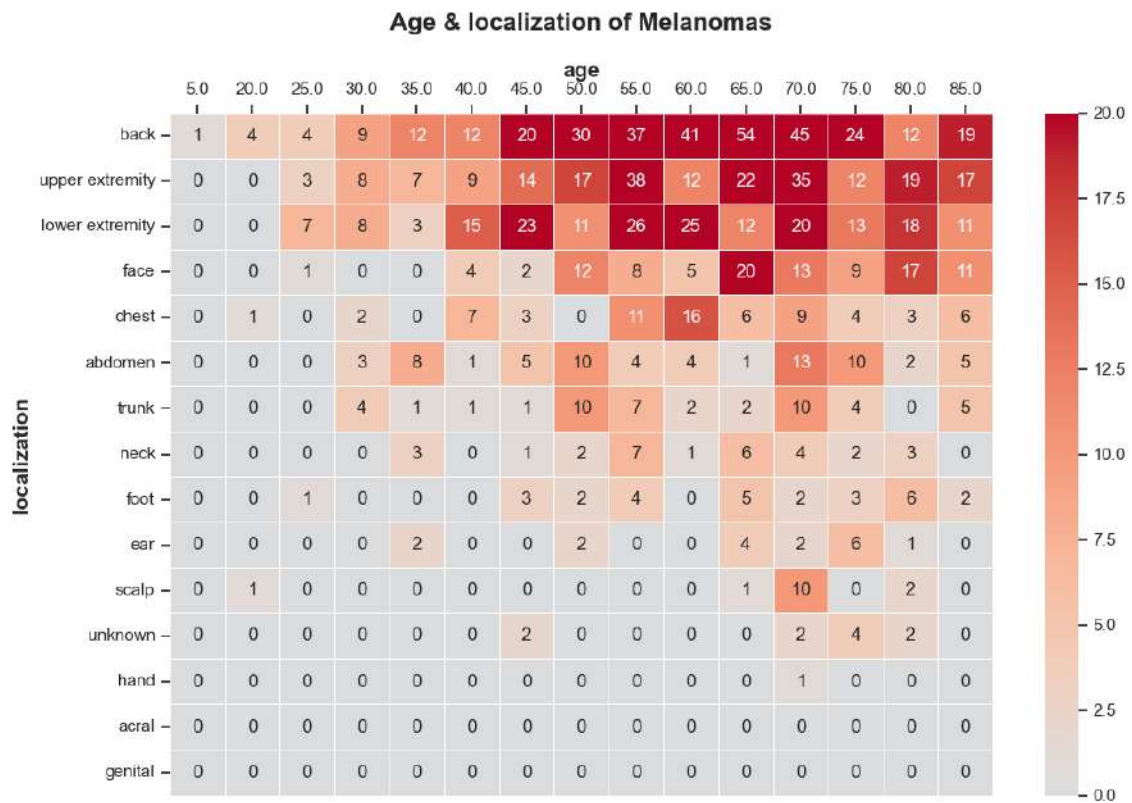


Figure 2: Zoom-in on melanoma in the dataset: number of images by age & localization

## 2.2 Data pre-processing

The pre-processing of skin lesion images was done by using Keras ImageDataGenerator. The images in the dataset were downsampled to 224X224 pixel resolution from 600X450 pixel resolution to make images compatible with the models described later. The 10015 images in the dataset were split into the training set (8516 images) and validation set (1499 images).

## 2.3 Imbalanced Data

When dealing with imbalanced data in clinical research images, there are various techniques that can be employed to improve the performance of classifiers. Some of the techniques used in this paper include:

### **2.3.1 Ensemble Techniques**

such as bagging, boosting, or stacking can be used to combine multiple models, where different models may perform better on different classes and improve the performance of the classifier. This can be useful in cases of imbalanced data.

### **2.3.2 Data Augmentation**

techniques can be used when the image dataset is small or imbalanced. It is a good practice to artificially introduce sample diversity by applying random yet realistic transformations to the training images, such as random horizontal flipping or small random rotations. This helps expose the model to different aspects of the training data while slowing down overfitting.

### **2.3.3 Transfer Learning**

Transfer learning [3] can be used to fine-tune a pre-trained model on a related task and adapt it to imbalanced data. This can help improve the performance of the classifier and reduce the amount of data required for training. As can be discerned from Figure 3, the architectural scheme of our project lays out the mechanism of the transfer learning process for skin lesion classification. This involves employing the power of pre-existing models, namely ResNet50, VGG19, Xception, and InceptionResNetV2, originally trained on vast datasets like ImageNet. The primary role of these pre-trained models in our project is to act as sophisticated feature extractors, adept at transferring learned knowledge to our specific, imbalanced skin lesion classification task. To fine-tune these models for our project, we modify their final layers, a step that essentially customizes them to our specific classification requirements. This process optimizes class recognition, including those classes that may be underrepresented in the dataset. In essence, this project leverages the efficacy of transfer learning in addressing real-world problems that present imbalanced data, thereby curtailing training duration while enhancing the performance of our classification model.

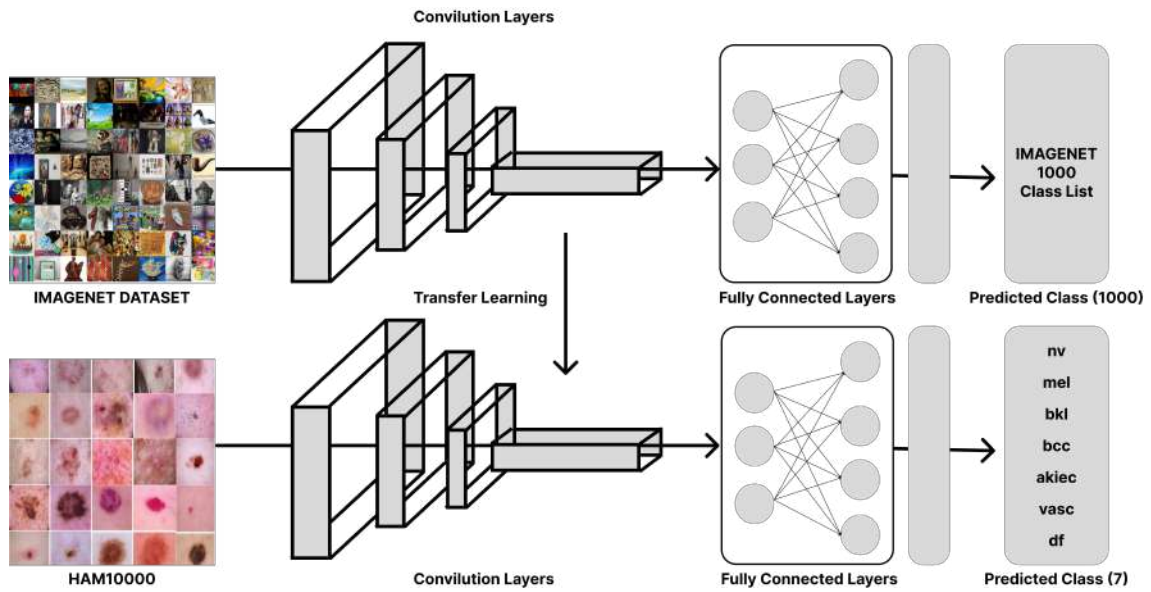


Figure 3: Transfer Learning from ImageNet

By using one or more of these techniques, it is possible to improve the performance of classifiers on imbalanced clinical research image data and ensure that the minority class is not overlooked. It is important to note that in clinical research, the accuracy of a classifier is not the only metric of interest. Other metrics such as sensitivity, specificity, and positive predictive value may be more relevant in certain contexts. For example, in a diagnostic setting, it may be more important to maximize sensitivity (i.e., correctly identify all individuals with a specific condition) at the expense of specificity (i.e., incorrectly identify some individuals without the condition as having it). Furthermore, it is important to carefully consider the potential consequences of misclassifying samples from the minority class. For example, misclassifying rare cancer as benign could have serious consequences for the patient. Overall, dealing with imbalanced data in clinical research images requires careful consideration of the potential consequences of misclassification, as well as the use of appropriate techniques to ensure that the minority class is not overlooked. By using a combination of resampling techniques, ensemble techniques, data augmentation, and different evaluation metrics, it is possible to improve the performance of classifiers on imbalanced data sets.



## 2.4 Data Augmentation

HAM10000 dataset has an unbalance distribution of images among the seven classes. Data Augmentation brings an opportunity to rebalance the classes in the dataset, alleviating other minority classes. Data Augmentation is an effective means to expand the size of training data by randomly modifying several parameters of training data images like rotation range, zoom range, horizontal and vertical flips, fill mode, zoom, shear, etc. [4]

The paper also suggests that color transforms (saturation, contrast, brightness) can be applied, however, they perform worse than the no augmentation scenario. One option for the worse results is that the color (pigment) of skin lesions is indicative to the lesion type upon analyzing it, and when we modify it we make it harder to differentiate between the types. Since we also received worse results with those augmentations, we dropped them from this report.

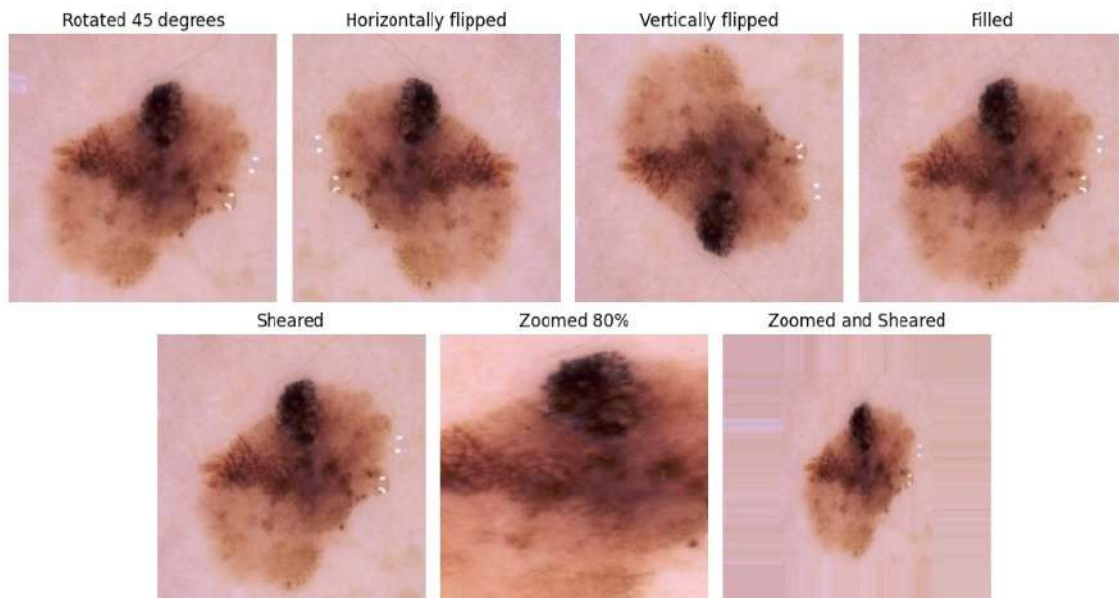


Figure 4: Image augmentation example

## 2.5 Training Algorithms

For our study, we utilized transfer learning with several popular pre-trained convolutional neural network (CNN) models, namely ResNet-50, VGG19, Xception, and InceptionV2, to achieve accurate classification of skin lesions. These models have been extensively used in computer vision tasks and have shown excellent performance in image recognition.

### **2.5.1 ResNet-50**

ResNet-50 is a deep CNN architecture that addresses the vanishing gradient problem through the use of residual connections. These connections allow for more efficient gradient propagation during training, enabling the network to capture intricate features and patterns in images. With its 50 layers, ResNet-50 is well-suited for complex visual recognition tasks.

### **2.5.2 VGG19**

VGG19 is a deep CNN model known for its simplicity and effectiveness. It consists of 19 layers, primarily utilizing 3x3 convolutional filters. The uniform architecture of VGG19 makes it easy to understand and implement. It excels at capturing fine-grained details in images, making it a popular choice for precise image analysis and classification tasks.

### **2.5.3 Xception**

Xception is a novel CNN architecture that introduces depthwise separable convolutions. Instead of performing standard convolutions, Xception separates the spatial and channel-wise filtering, resulting in more efficient feature extraction. This design significantly reduces the number of parameters and computational complexity while maintaining high performance. Xception is particularly suitable for resource-constrained scenarios, such as mobile and embedded vision applications.

### **2.5.4 InceptionResNetV2**

Combines the Inception and ResNet architectures, integrating both the multi-scale feature extraction capabilities of Inception and the residual connections of ResNet. It features a deep network with inception modules, residual connections, and auxiliary classifiers. InceptionResNetV2 excels at capturing complex features and has shown strong performance in various image classification tasks.

These models were initially trained on large-scale datasets, in particular on the ImageNet Challenge, which contains millions of images spanning various object classes. Leveraging transfer learning, we fine-tuned these pre-trained models on our skin lesions dataset. Fine-tuning involved retraining the models on our specific task with a small number of epochs and updating

layers’ weights accordingly. During training, we employed a batch size of 64 and trained the models for 30 epochs. The training process utilized the Categorical Crossentropy loss function, Adam optimizer, and the Accuracy evaluation metric. This metric provided insights into the models’ ability to correctly classify the lesions and rank them based on confidence scores. By leveraging the capabilities of these pre-trained CNN models through transfer learning, our aim was to achieve robust and accurate classification of skin lesions. Ultimately, this contributes to the early detection and diagnosis of skin conditions, improving patient outcomes in dermatology.

This is also the appropriate section to mention the current state-of-the-art in image classification, known as Vision Transformers (ViT). ViT has surpassed ResNet, which was previously considered the leading CNN model before the emergence of ViT. However, when trained on mid-sized datasets like ImageNet (or in our case, HAM10000) without robust regularization, ViT achieves slightly lower accuracies compared to comparable-sized ResNets. If the dataset from training is sufficiently large (at least 100 million images), ViT beats CNNs by a small margin. Consequently, for the purposes of this study, we will focus on the realm of CNNs and not utilize ViT [5].

Model	Color	Augmentation	Accuracy	Precision	Recall	F1
ViT32	True	True	0.85	0.71	0.80	0.74
ViT32	False	True	0.71	0.49	0.67	0.53

Table 2: ViT-Models metrics result

## 2.6 Evaluation metrics

The overall performance of the model was evaluated using various evaluation metrics, including Accuracy, Macro Average Precision (MAP), Macro Average Recall (MAR), Macro Average F1-score (MAF). We use macro average scores when we need to treat all classes equally to evaluate the overall performance of the classifier against the most common class labels. Recall, Precision, and F1-score are computed using the following formulas:

**Accuracy** is calculated as the ratio of the sum of true positives (TP) and true negatives (TN) to the sum of TP, TN, false positives (FP), and false negatives (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

**Precision** measures the proportion of correctly identified positive instances (TP) out of the sum of TP and false positives (FP).

$$Precision = \frac{TP}{TP + FP} \quad (2.1)$$

$$MacroAvgPrecision = \frac{1}{n} \sum_{i=1}^n Precision_i \quad (2.2)$$

**Recall** also known as sensitivity or true positive rate, calculates the proportion of correctly identified positive instances (TP) out of the sum of TP and false negatives (FN).

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (3.1)$$

$$MacroAvgRecall = \frac{1}{n} \sum_{i=1}^n Recall_i \quad (3.2)$$

**F1** is the harmonic mean of precision and recall, providing a balanced measure of their performance. It is calculated using the formula:

$$F1 = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \quad (4.1)$$

$$MacroF1 = \frac{1}{n} \sum_{i=1}^n F1_i \quad (4.2)$$

These metrics enable us to comprehensively evaluate the accuracy, precision, recall, and overall performance of the model in classifying skin lesions.

### 3 Results

Model Evaluation was performed by calculating the F1-score, classification report, and confusion matrix. Further, the loss and accuracy curves were plotted to validate the model's performance for the optimization and prediction phase.

### 3.1 Model Validation

	Model	Color	Augmentation	Localization	Accuracy	Precision	Recall	F1
1	<b>ResNet50</b>	<b>RGB</b>	<b>True</b>	<b>All</b>	<b>0.89</b>	<b>0.77</b>	<b>0.81</b>	<b>0.79</b>
2	<b>ResNet50</b>	Grayscale	True	All	0.81	0.63	0.74	0.67
3	<b>Xception</b>	RGB	True	All	0.87	0.75	0.82	0.78
4	<b>Xception</b>	Grayscale	True	All	0.80	0.61	0.71	0.66
5	<b>InceptionResNetV2</b>	RGB	True	All	0.85	0.74	0.78	0.76
6	<b>InceptionResNetV2</b>	Grayscale	True	All	0.83	0.67	0.73	0.70
7	<b>VGG19</b>	RGB	True	All	0.73	0.57	0.69	0.61
8	<b>ResNet50</b>	RGB	False	All	0.87	0.77	0.78	0.76
9	<b>ResNet50</b>	RGB	True	Lower extremity	0.88	0.80	0.79	0.79

Table 3: Models metrics result

### 3.2 Confusion Matrix

The model’s confusion matrix, spanning seven classes, evaluates the True and Predicted labels for each image in the validation set. The model displayed remarkable proficiency in classifying Melanocytic nevi (nv), accurately predicting 945 out of the 1005 instances (based on the recall of 0.94). It also demonstrated high precision with Basal cell carcinoma (bcc) and Melanoma (mel), accurately predicting 67 out of 77 instances (recall of 0.87) and 111 out of 166 instances (recall of 0.67), respectively. The model had some difficulty in diagnosing Benign keratosis-like lesions (bkl) due to their visual similarity to Melanoma and Melanocytic nevi, correctly predicting 131 out of 164 instances (recall of 0.80).

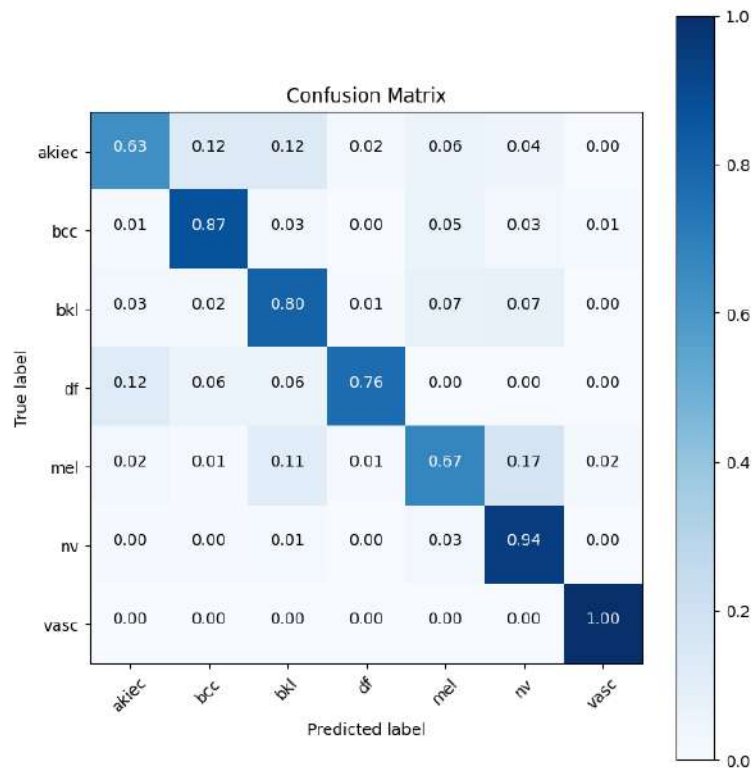


Figure 5: Confusion Matrix of the best result (ResNet50, RGB with augmentations)

Overall, while the model showed solid performance in identifying some skin conditions, there is room for improvement in its ability to accurately classify others.

### 3.3 Loss and accuracy curves

Over 30 epochs, the model exhibited consistent improvement in both the training and validation phases. The training commenced with a loss of 5.700743 and an accuracy of 52.10%. Simultaneously, the validation loss was 5.401198 with an accuracy of 75.32%. With training progression, the model demonstrated a constant decline in losses and a steady increase in accuracy. By the 10th epoch, the model had reduced the training loss to 2.258751 and increased accuracy to 83.22%. Concurrently, the validation loss and accuracy improved to 2.410881 and 82.79%, respectively. By the final 30th epoch, the model managed to significantly reduce the training loss to 1.052586, and the accuracy reached an impressive 93.15%. Similarly, the validation loss and accuracy improved to 1.359893 and 89.19%, respectively.

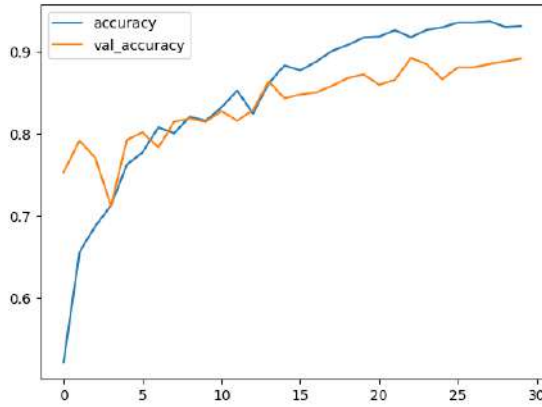


Figure 6: Accuracy vs. epoch

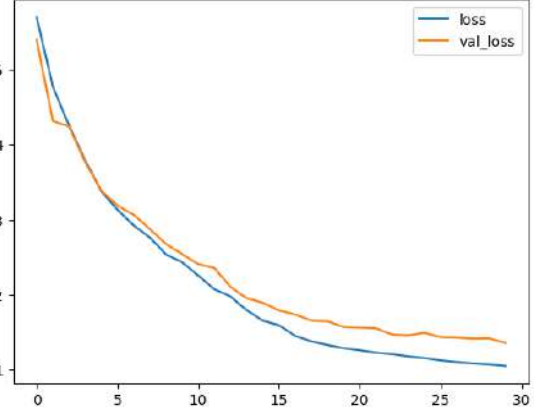


Figure 7: Loss vs. epoch

In summary, the model consistently improved in learning capability, demonstrated by declining losses and increasing accuracy across the epochs. It performed well on unseen data, indicating good generalization ability. Mild overfitting was observed as the model’s training accuracy was consistently higher than the validation accuracy.

### 3.4 Grad-CAM

Grad-CAM[6], or Gradient-weighted Class Activation Mapping, is a technique for visualizing the parts of an image that are most relevant to a model’s prediction. It works by taking the gradients of the model’s output with respect to the input image, and then weighting these gradients by the activations of the last convolutional layer. The resulting heatmap can then be overlaid on the original image to show which parts of the image are most important for the model’s prediction. Grad-CAM is a useful tool for understanding how deep learning models work. It can be used to identify the features that a model is using to make its predictions, and to see how these features change when the model is presented with different images. Grad-CAM can also be used to debug models, and to identify areas where the model is making mistakes.

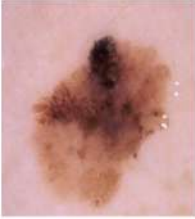
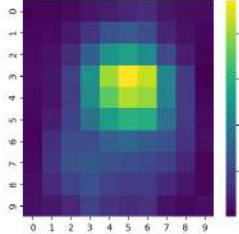

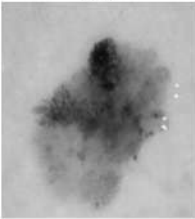
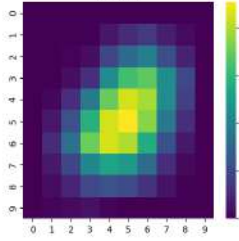
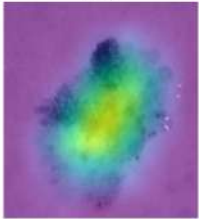

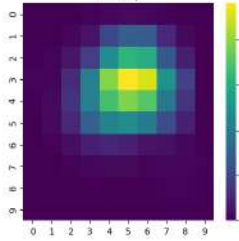

#	Model	Color	Augmentation	Output		
1	ResNet50	RGB	True			
2	ResNet50	Gray	True			
9	ResNet50	RGB	False			

Table 4: Grad-CAM (ResNet50) output for mel

we observed that the model trained on RGB images exhibited a stronger emphasis on color and texture when focusing on skin lesions. This indicates that the model relied on these visual features to make predictions. On the other hand, when the model was trained on grayscale images, it exhibited a greater focus on the shape of the skin lesions. This suggests that the model placed more importance on the structural characteristics and overall shape of the lesions rather than the color and texture details. Overall, these findings suggest that the choice of image type during training influenced the model's attention towards specific visual features when making predictions about skin lesions.




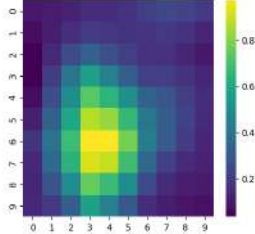

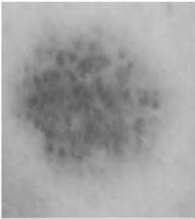
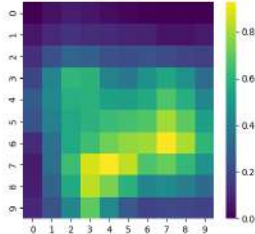
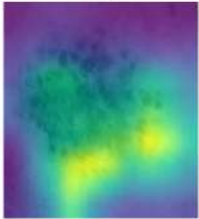

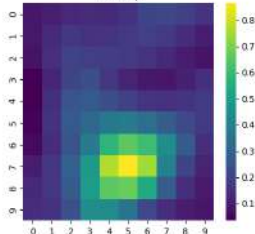
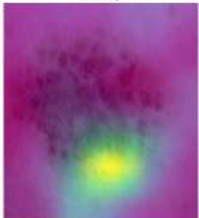
#	Model	Color	Augmentation	Output		
1	ResNet50	RGB	True			
2	ResNet50	Gray	True			
9	ResNet50	RGB	False			

Table 5: Grad-CAM (ResNet50) output for nv

Using the Grad-CAM for activation map visualization offers simplicity, interpretability, versatility, and fine-grained results. It allows us to visualize the regions of an image that are most influential in determining the model’s prediction for a specific class, providing valuable insights into the model’s decision process and localizing important features within an image. Leveraging Grad-CAM for each class empowers us to make informed decisions for model improvement, refining the model’s architecture, hyperparameters, or training strategies to enhance its capabilities and achieve higher performance levels. We can plot the mistakes made by the models, such as classifications between "akiec" vs. "bkl" and "nv" vs. "mel". This analysis provides a deeper understanding of how the model makes decisions and highlights areas for improvement, which will be discussed in more detail in section 3.5.4.

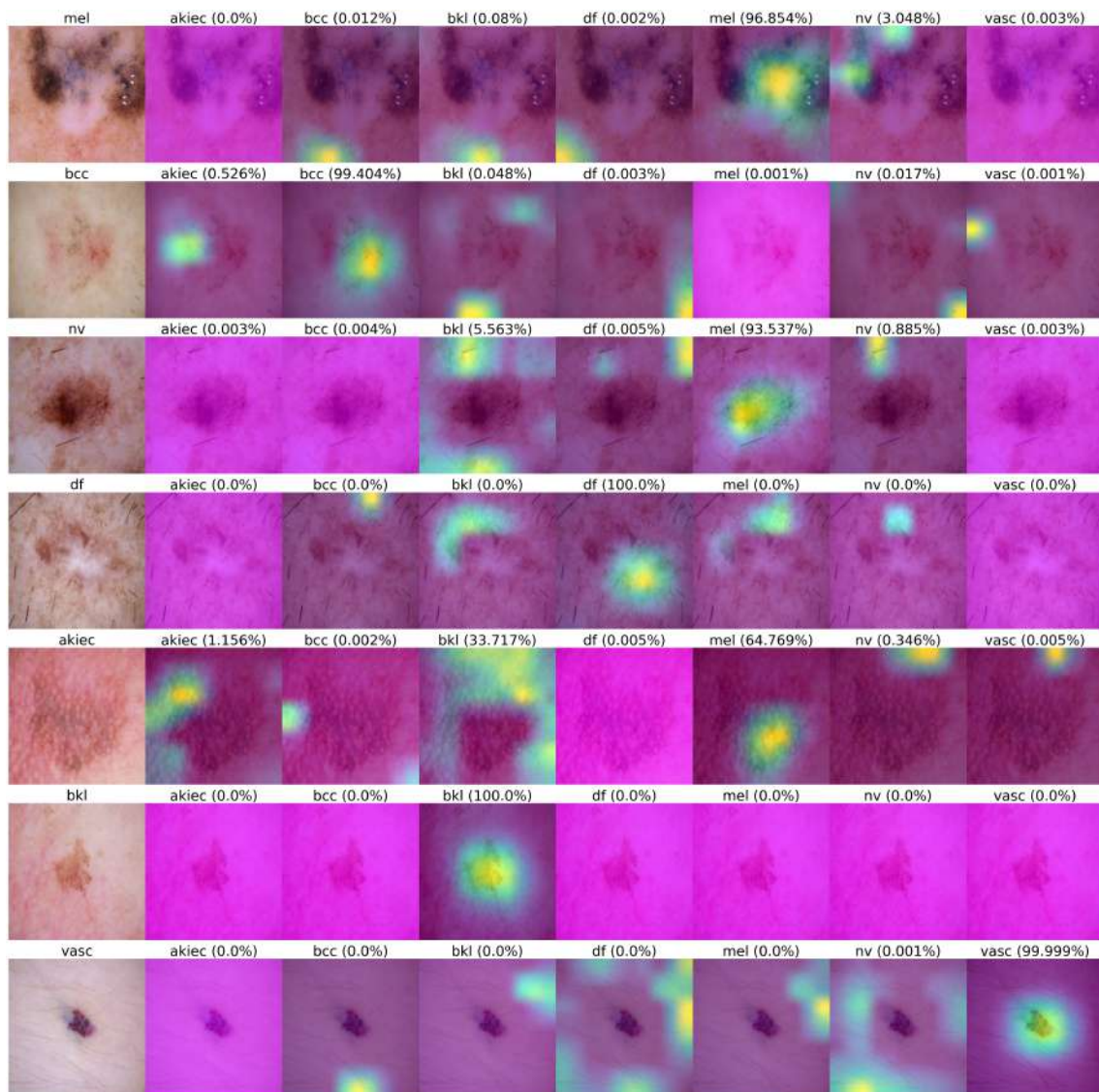


Figure 8: Activation map for each class: ResNet50-RGB

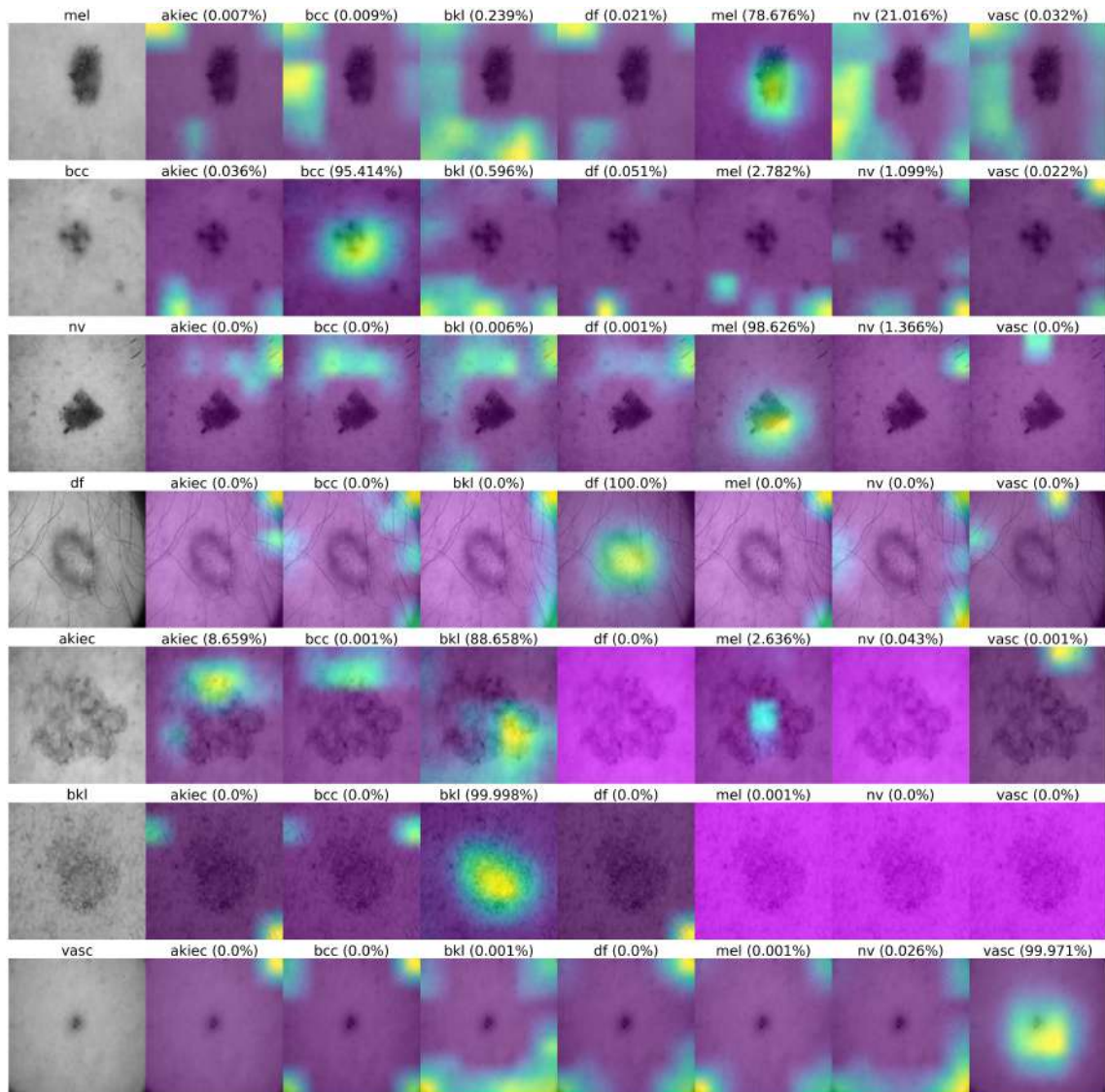


Figure 9: Activation map for each class: ResNet50-Grayscale

## 3.5 Ensemble Methods

### 3.5.1 Technique 1: Averaging

This is a basic form of model averaging ensemble technique. In this approach, multiple models are trained independently on the same dataset, and their predictions are averaged to generate the final prediction. In this particular implementation, we load multiple pretrained CNN models and predict the labels of the testing images using each model. The predictions of each model

are then averaged to generate the final prediction for each image. By averaging the predictions of multiple models, we hope to reduce the variance in the predictions and improve the overall accuracy of the ensemble.

### 3.5.2 Technique 2: Voting

The ensemble technique used above is called voting. Voting is a simple but effective ensemble technique that combines the predictions of multiple models by taking a majority vote. In the example notebook, the predictions of three CNN models are combined to produce a single prediction. The model with the highest prediction score for each image is selected as the ensemble prediction. Voting is a robust ensemble technique that can be used with a variety of machine learning models. It is particularly effective for reducing the variance of a model, which can lead to improved accuracy. Tie-breaking is solved by sorting labels in ascending order and selecting the first label. Here are some of the advantages of using voting as an ensemble technique:

- It is simple to implement.
- It is effective for reducing variance.
- It can be used with a variety of machine learning models.

Here are some of the disadvantages of using voting as an ensemble technique:

- It may not improve accuracy as much as other ensemble techniques.
- It can be computationally expensive to train multiple models.

Overall, voting is a simple and effective ensemble technique that can be used to improve the accuracy of machine learning models.

### 3.5.3 Ensemble results

In our case, we chose to make an ensemble using the top-3 RGB models: ResNet50, Xception, and InceptionResNetV2 to gain a better result than the results of the individual models: Using technique 1 we get  $F1 = 0.82$  compared to 0.79, 0.78, and 0.76 of the models in the ensemble (an improvement of 3% over the best performing model). A substantial improvement in performance.

In terms of technique 2, the ensemble result is  $F1 = 0.81$  (an improvement of 2% over the best performing model).

These improvements over the best performing model indicates that the ensemble model is able to leverage the strengths of the individual models and make more accurate predictions. This improvement suggests that the ensemble is effectively capturing diverse patterns and making more robust predictions compared to any single model alone.

\* Improvements in final results are seen also when considering the corresponding grayscale images, although with smaller values in percentages.

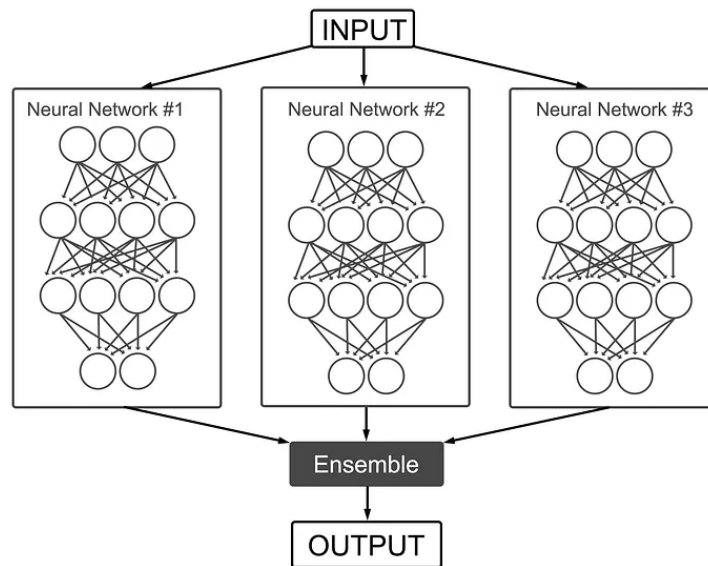


Figure 10: Ensemble high-level illustration

#### 3.5.4 Mix-Model

We wanted to add another layer of analysis to our results, and realized that our best model's weak-point is differentiating between Melanoma and Melanocytic nevi lesions - 17% of times the model is presented a Melanoma lesion, it falsely predicts Melanocytic nevi. The mixed model approach that we implemented combines the strengths of two different machine learning models, specifically a generally best model and a second, more specialized model.

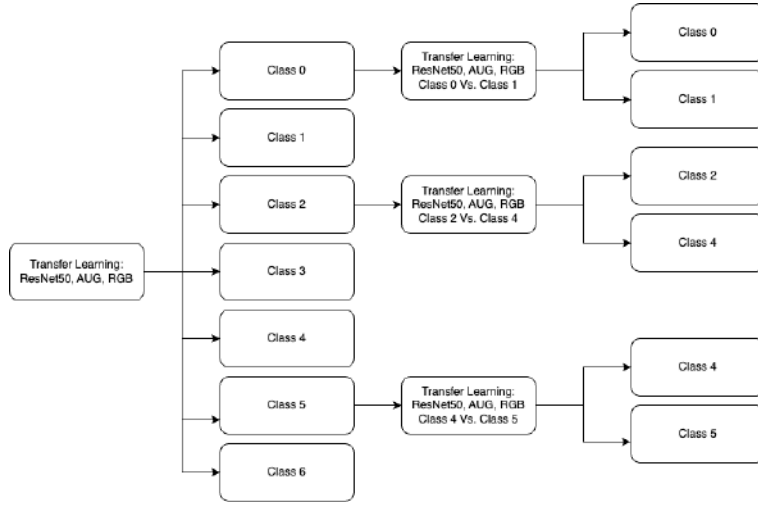


Figure 11: Mix-Model illustration

In order to prevent the model from benefiting from knowledge of the test set during training, we will use the same training data that was provided to the base model as the training data for the mix model. This ensures that the mix model is not exposed to any information from the test set and helps maintain the integrity of the evaluation process. By using the same training data, we can evaluate the performance of the mix model in a fair and unbiased manner. This approach ensures that the mix model is tested on unseen examples and provides a reliable assessment of its generalization capabilities.

**The base (general) model:** The base model is based on transfer learning using the ResNet50 architecture [ 3.1] and was trained on RGB images with augmentations, which consists of images from 7 different classes. This is the best model from Table 2.

**The specialized models:** The mixed model approach in this project consists of four models, each addressing specific weaknesses or challenges of the base model in differentiating between certain types of skin lesions.

The base model utilizes the ResNet50 architecture and is trained on augmented RGB images, providing a general classification capability across a wide range of skin lesion classes. However, it may struggle with certain challenging differentiations.

**mel vs. nv:** The second specialized model focuses on distinguishing between Melanoma and Melanocytic nevi. These two classes pose difficulties for the base model due to their visual similarities and overlapping features. By training a model specifically on this classification task, it can develop a deeper understanding of the nuanced differences, improving its accuracy in distinguishing between Melanoma and Melanocytic nevi.

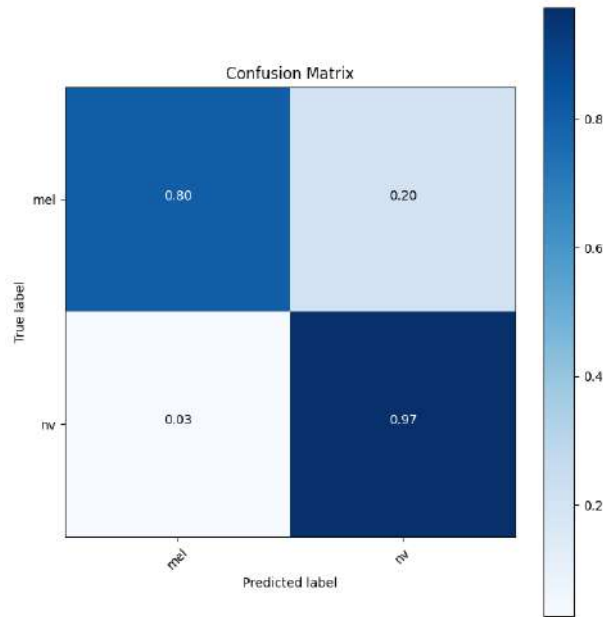


Figure 12: confusion matrix: mel Vs. nv

**akiec vs. bcc:** The third specialized model is designed to differentiate between akiec (Actinic keratoses and intraepithelial carcinoma) and bcc (Basal cell carcinoma) lesions. The base model might face challenges in accurately classifying these two classes, as they can share certain visual characteristics such as redness, scaling, and ulceration. By training a specialized model exclusively on this differentiation, it can learn to identify subtle features that distinguish akiec from bcc lesions, enhancing its accuracy in this domain.



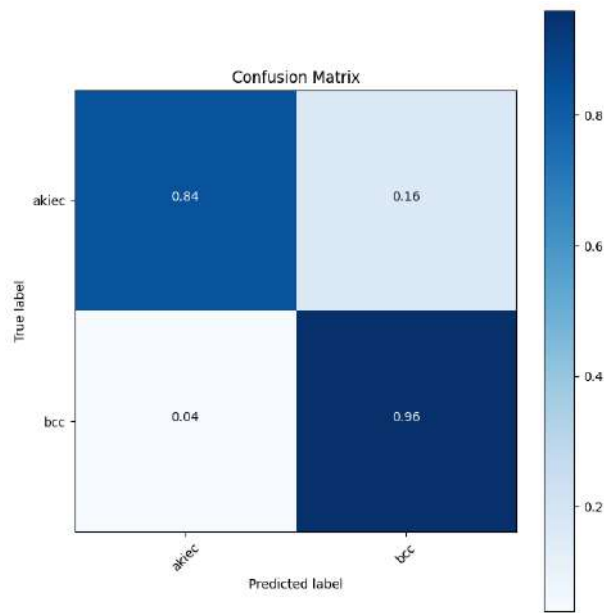


Figure 13: confusion matrix: akiec Vs. bcc

**mel vs. bkl:** The fourth specialized model is focused on the classification of Melanoma versus Benign keratosis-like lesions (bkl). These two classes can exhibit similar visual features, including irregular borders, pigment variations, and asymmetric patterns. The base model's ability to accurately differentiate between Melanoma and bkl might be limited. By training a dedicated model for this specific comparison, it can gain expertise in discriminating between the unique characteristics of Melanoma and bkl lesions, thereby improving classification accuracy.



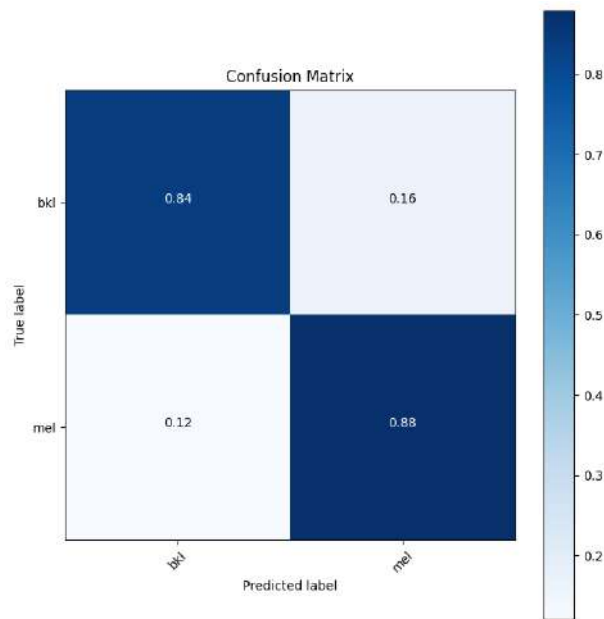


Figure 14: confusion matrix: mel Vs. bkl

With the mixed model approach, the general model serves as the initial classifier. If it predicts a skin lesion as Melanoma, the specialized models for Melanoma versus Melanocytic nevi, akiec versus bcc, and Melanoma versus bkl are selectively employed to refine the predictions. This selective utilization of specialized models addresses the weaknesses of the base model and enhances the accuracy and reliability of the overall classification process.

The use of multiple specialized models in the mixed model approach allows for a more precise and accurate classification of diverse skin lesion types present in the HAM10000 dataset. By focusing on challenging differentiations, the approach aims to assist doctors in diagnosing skin cancer more accurately and efficiently.

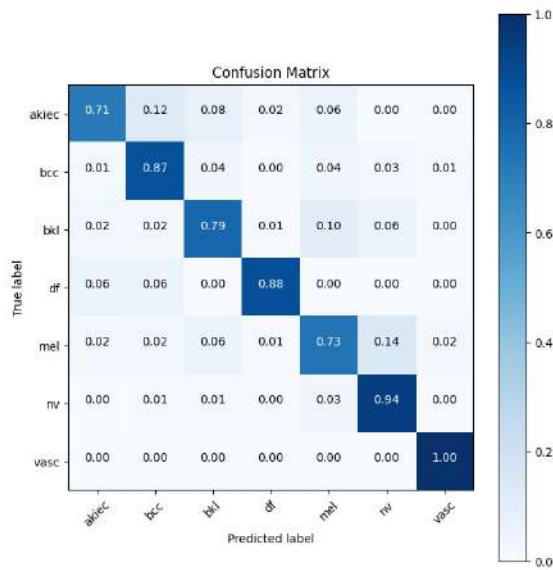


Figure 15: Confusion Matrix: Mix-Model

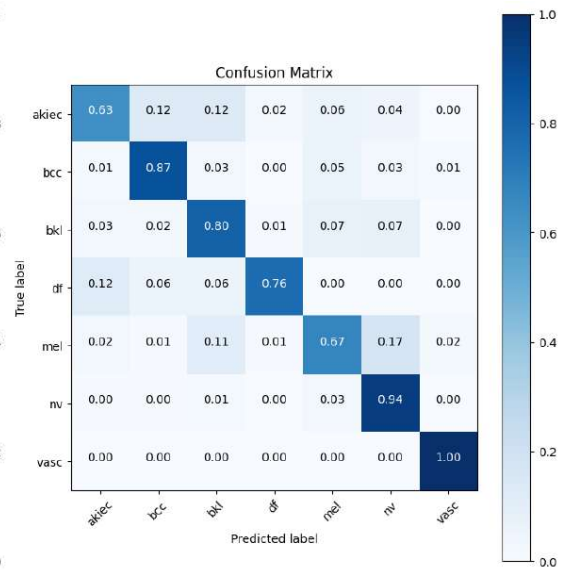


Figure 16: Confusion Matrix: Base-Model

Model	Accuracy	Precision	Recall	F1
ResNet50 (RGB with augmentations)	0.88	0.77	0.81	0.79
<b>Mix-Model</b>	<b>0.89</b>	<b>0.78</b>	<b>0.85</b>	<b>0.81</b>

Table 6: Mix-Model Vs. Resnet50 (Aug,RGB) metrics result

The ResNet50 model achieved an accuracy of 0.88, a precision of 0.77, a recall of 0.81, and an F1 score of 0.79. However, the Mix-Model, which is highlighted, has demonstrated superior performance across all metrics compared to the ResNet50 model. It achieved an accuracy of 0.89, a precision of 0.78, a recall of 0.85, and an F1 score of 0.81, and shows significant improvements in certain classes over the ResNet50 model. In particular, it shows improved precision for the akiec classes, and improved recall for the bcc class. Notably, the precision and recall for the nv class remain the same across both models, indicating that both models perform exceptionally well for this class. Despite the improvements in certain classes, the f1-score for the mel class remains the same in both models, indicating that there is room for improvement in the prediction of this particular class for the Mix-Model.

## 4 Conclusions & Discussions

Based on the results obtained from the ham10000 dataset using transfer learning, it can be observed that augmentations have a positive impact on the performance of the model. When all the data, including augmentations, was used for training, an F1-score was achieved. This indicates that the model was able to achieve a good balance between precision and recall, resulting in a reliable performance.

Furthermore, it is interesting to note that when training the model using only images from a specific area ("lower extremity"), with augmentations, the same F1-score was achieved. This suggests that the model was able to generalize well and capture the important features and patterns related to skin cancer classification specific to that area. It implies that the transfer learning approach effectively leveraged knowledge gained from training on a broader dataset to perform well on a localized dataset.

On the other hand, the worst results were obtained when training the model on all the data without any augmentations. This implies that augmentations play a crucial role in enhancing the model's ability to learn and generalize from the data. Augmentations introduce variations and diversify the training samples, allowing the model to better handle different scenarios and improve its overall performance.

In terms of further improving the final predictions: various ensemble techniques can be used as we showed (Averaging, Voting, etc.).

Moreover, specialized models can be utilized to re-validate some predictions given by the general model for specific lesion types (Mix-Model approach). The Mixed-Model approach leverages the strengths of the first model, trained on a broad range of classes, and the specialization of the second, third and fourth models to improve accuracy for specific classes of interest. It is an innovative strategy that combines the power of different models to achieve better overall performance in classification tasks.

In terms of explainability, visualizations such as Grad-CAM can be used to infer how the model "sees" each prediction.

## 5 References

- [1] Harald Kittler et al. *Dermatoscopy: Pattern Analysis of Pigmented and Non Pigmented Lesions*. 2nd ed. facultas.wuv Universitäts, 2016. ISBN: 9783708913858. URL: [https://books.google.co.il/books/about/Dermatoscopy.html?id=W1I-jwEACAAJ&redir\\_esc=y](https://books.google.co.il/books/about/Dermatoscopy.html?id=W1I-jwEACAAJ&redir_esc=y).
- [2] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. *HAM10000: a large collection of multi-source dermatoscopic images of common pigmented skin lesions*. Version V1. 2018. DOI: 10.7910/DVN/DBW86T. URL: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>.
- [3] Wikipedia. *Transfer learning* — *Wikipedia, The Free Encyclopedia*. [Online; accessed 22-May-2023]. 2023. URL: [https://en.wikipedia.org/wiki/Transfer\\_learning](https://en.wikipedia.org/wiki/Transfer_learning).
- [4] F. Perez et al. “Data Augmentation for Skin Lesion Analysis”. In: *ArXiv*. 2018. DOI: 10.1007/978-3-030-01201-4\_33.
- [5] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].
- [6] Ramprasaath R. Selvaraju et al. “Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization”. In: *CoRR* abs/1610.02391 (2016). arXiv: 1610.02391. URL: <http://arxiv.org/abs/1610.02391>.