Skin Cancer Classifier using Dermatoscopic Images of Pigmented Lesions

Final Report

Meha Gupta
Samarth Sinha
Ines Bosch-Alfonso
Valentina Manferrari

Word Count: 2433
Penalty: 0

1. Introduction

Skin cancer is the 19th most common form of cancer for humans [1]. As with most forms of cancer, early detection and treatment of skin cancer can have life-altering results. Skin cancer is diagnosed through a skin biopsy that involves a careful examination of the patient's moles, skin lesions and overall skin tone [2]. This process can be lengthy, costly and complex, and it typically requires a high level of technical skill and experience.

Our project goal is to design a classifier that is able to classify dermatoscopic images of skin lesions as being malignant or benign. Machine learning would be of appropriate use in this field as a classifier trained with images of skin cancer lesions would greatly aid in this early detection. Our design will allow those without formal medical training to easily determine if a skin lesion/mole is cancerous by simply providing an image of the lesion.

2. Illustration

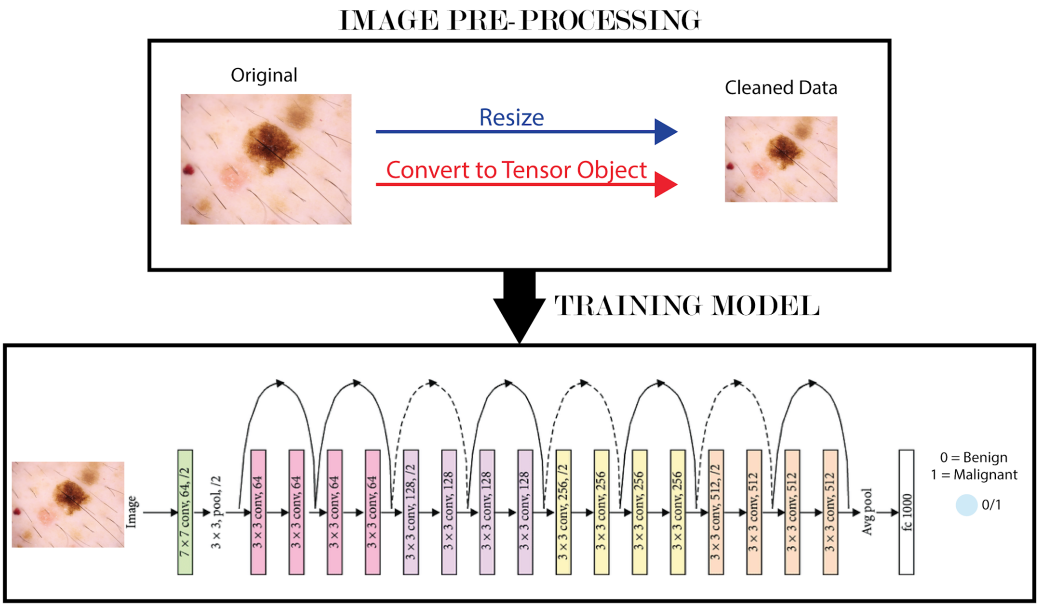

Figure 1. A diagram visualizing our final ResNet18 model

3. Background & Related Work

Following the advancements made in computer vision and the introduction of publicly-available large dermoscopic datasets with different types of benign and malignant skin lesions, the use of dermoscopic images in skin cancer classification algorithms has become a popular field of research. In a landmark study, Esteva et al. [3] compared the performance of a CNN model (GoogleNet Inception-v3 Architecture), trained on a combined dataset of 129,450 clinical and dermatological images (representing 2032 different skin lesions), with the performance of 21 board-certified dermatologists for skin cancer

classification across three categories, and successfully demonstrated that the performance of the neural network was at par with dermatologist performance.

In another investigation, Gouda et al. [4] trained a ResNet-34 model over the preprocessed ISIC 2019 Challenge dataset (which includes 25,331 dermatoscopic images of skin lesions and moles from 8 different classes of possible skin disease like Melanoma, Basal cell carcinoma, Actinic Keratosis etc.) to build a multi-class skin disease classifier that achieved an overall 92% accuracy on novel test images.

4. Data Processing

Our original data source has been the ISIC 2020 Challenge dataset which includes images and labels for examples of both malignant and benign skin lesions [5]. It contained 33,126 dermoscopic images of over 2000 different patients; the images were all .jpg and were all sized 1024x1024 pixels. However, although the size of this database might have been beneficial for the model training process, it was too big to manage on Google Colab. Therefore we had to select a new dataset from Kaggle [6]: the new images were taken from the same source as our original one but the overall size of the database was smaller, containing only 3,297 images. Figure 2 below shows a representative sample of our dataset.

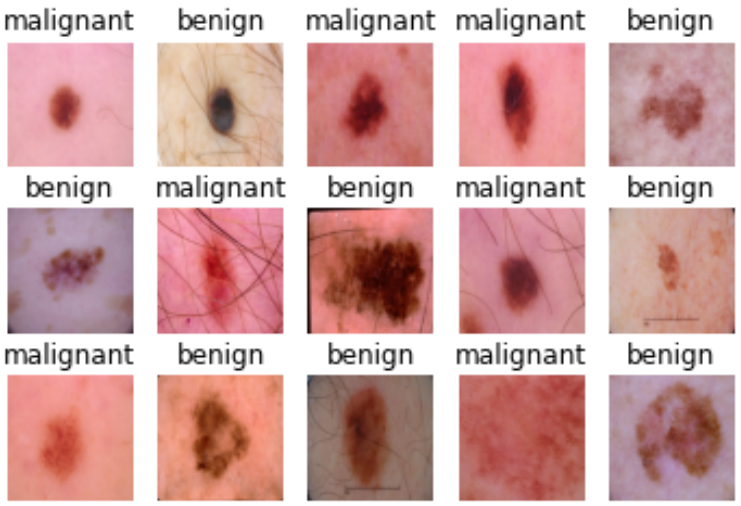

Figure 2. 15 sample images from our final dataset

Our data processing consisted in resizing the images to 224x224 pixels and converting them to tensors. We also checked that the dataset was balanced, since we calculated an almost equal amount of malignant and benign labels (1497 vs 1800) the team agreed that the dataset was ready for training. We proceeded in splitting the skin images into training, validation and testing datasets with a proportion of 70:10:20 to get:

```
Training Samples: 2373
Validation Samples: 264
Testing Samples: 660
```

5. Architecture

We selected the ResNet 18 CNN architecture as our primary model. We trained over 8 models (including AlexNet, ResNet 50, ResNet 101 etc.) and made our final choice based on model performance on the validation set. As shown in Figure 3(a), the ResNet 18 model has five convolutional blocks. The first convolutional block (conv1) is composed of one convolutional layer, and the other convolutional blocks (conv2-conv5) are composed of two residual blocks each. As shown in Figure 3(b), each residual block is composed of two convolutional layers.

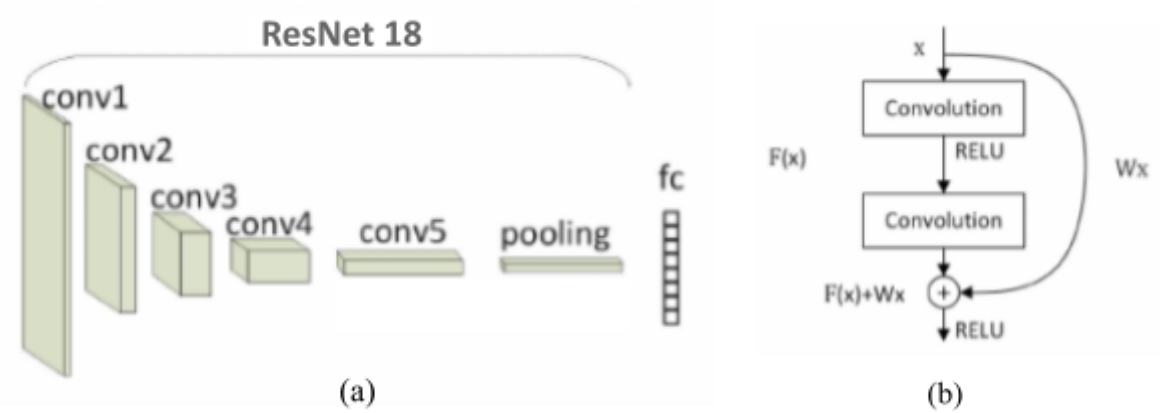

Figure 3. (a) ResNet 18 Architecture (b) Residual Block

We used the pretrained ResNet 18 Model from the PyTorch library using the following code:
with the following hyperparameters:

- 30 epochs
- Batch size of 128
- Learning rate of 0.00001

Training our model with these hyperparameters gave a reasonably high accuracy on the validation set and avoided any overfitting issue.

The final model can be viewed at this link.

6. Baseline Model

We selected a deep CNN as our baseline model. This model was chosen as our team wanted to have a similar architecture to compare our final model to. The CNN has 2 convolutional layers, 1 pooling layer, and 3 fully connected layers. The CNN model was based upon Google's LeNet model. The baseline model was created with the following code snippet:

```
class CNN(nn.Module):
```

```
    def __init__(self):
        super(CNN, self).__init__()
        self.name = "cnn"
           self.conv1 = nn.Conv2d(in_channels = 3, out_channels = 5,
kernel_size = 5, padding = 1)
        self.pool = nn.MaxPool2d(kernel_size = 2, stride = 2)
          self.conv2 = nn.Conv2d(in_channels = 5, out_channels = 10,
kernel_size = 5, padding = 1)
        self.fc1 = nn.Linear(10*54*54, 120)
        self.fc2 = nn.Linear(120, 84)
        self.fc3 = nn.Linear(84,2) # malignant or benign

    def forward(self, x):
        x = self.pool(F.relu(self.conv1(x)))
        x = self.pool(F.relu(self.conv2(x)))
        x = x.view(-1, 10*54*54)
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return x
```

The baseline model was then trained with the following hyperparameters:

- Adam optimizer
- 50 epochs
- Batch size of 32
- Learning rate of 0.001

With the above hyperparameters, our baseline model achieved the best results of:

- `Final Training Accuracy: 1.0`
- `Final Validation Accuracy: 0.8409090909090909`
- `Test Accuracy : 0.8424242424242424`

And the following training curves:

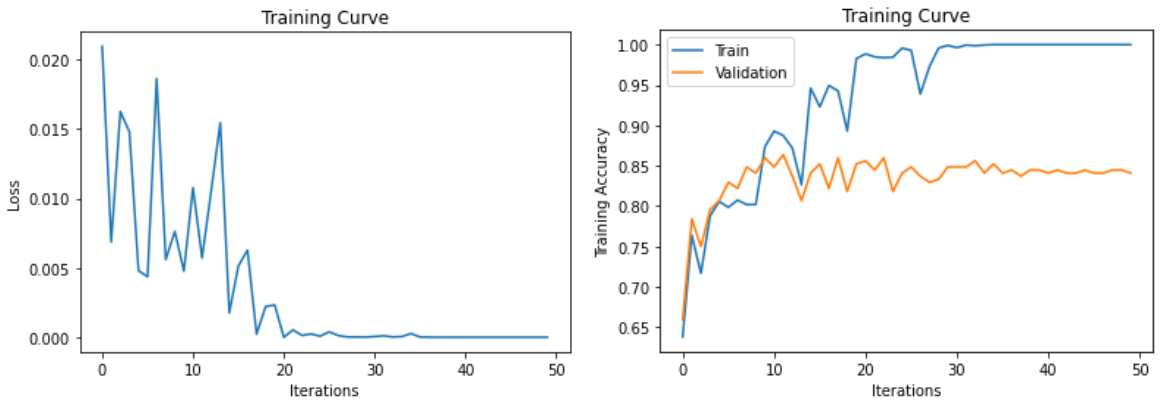

Figure 4. The resultant training loss (left) and accuracy (right) curves of the baseline model

Since we had an even split in our dataset amongst malignant and benign samples, we concluded that using accuracy was a suitable comparison metric for our final model and our baseline model.

Furthermore, our baseline model was tested on an unseen dataset of malignant skin lesions on people of colour (POC) which we created manually by searching through various dermatological journals [7][8][9][10]. This was done so that we could compare our final model's results on unseen data to our baseline model's results and thus get a more accurate picture of how well our final model was performing. As can be seen in figure 5 below, the baseline model performed quite poorly on samples of malignant skin lesions on POC, accurately classifying only 1/9 samples.

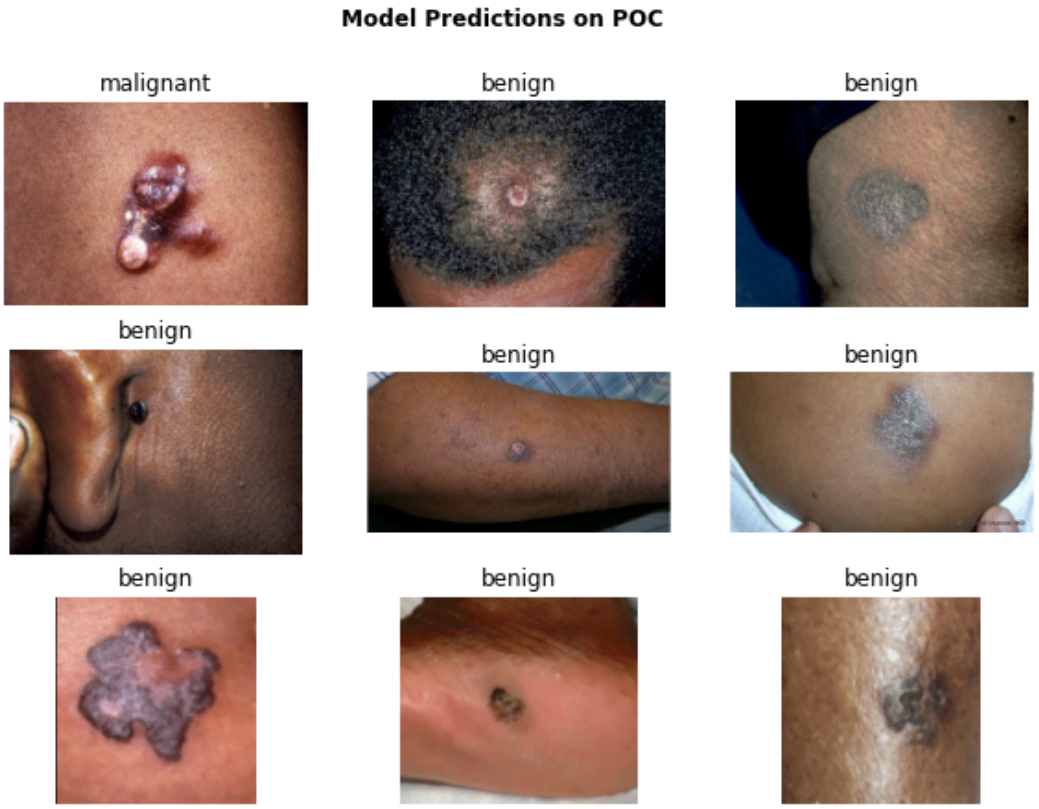

Figure 5. Baseline model predictions on malignant skin lesions on people of colours (POC)

The baseline model can be viewed here.

7. Quantitative Results

Model	# of layers	# of parameters	Best val. accuracy
ResNet18	50	11.7 M	**92.0%**
ResNet 50	50	25.5 M	91.6%
ResNet 101	101	44.5 M	90.5%
ResNet 152	152	60.2 M	88.6%
ResNext 50	50	25.0 M	90.5%
DenseNet 121	121	8.0 M	92.4%
DenseNet 161	161	28.7 M	90.9%

Table 1. Full benchmark results of 7 different state-of-the-art CV architectures. Each of the variants utilize ImageNet pretraining

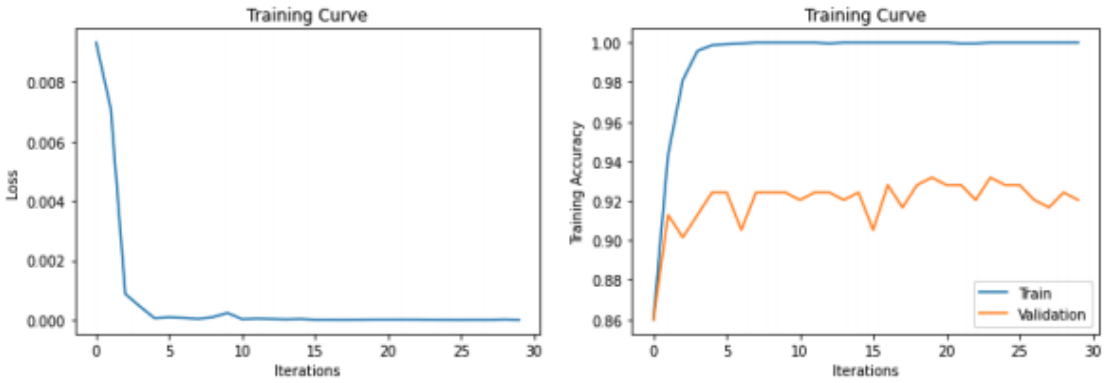

Figure 6: Training and validation accuracy curves

We exhaustively benchmark 7 different popular neural-network architectures and their variants. The main common theme between each of the variants is that they utilize residual or dense connections to propagate information between layers for inference. We also investigate the effect of scaling up the models and increasing the number of layers and its effect on validation performance, as larger models typically have better generalization properties when trained on large-scale datasets. Since the datasets are not very large for modern deep learning standards, we make use of ImageNet pretrained weights.

The best validation accuracy and the number of model parameters in each of the models are tabulated in Table 1, where we present results for: ResNet 18, 50, 101, 152, ResNeXt 50, and DenseNet 121, 161 [11][12][13]. We clearly see that the ResNet 18 has the highest validation accuracy compared to the other baseline models. Interestingly, ResNet 18 is also the shallowest model and has the second-fewest number of parameters, out of the different

variants tried. Since the training dataset is small, it is possible that very deep models, such as ResNet 152, have too many parameters and are unable to fine tune without memorizing the dataset. We finally test the model with the best performing validation accuracy on the test split of the dataset. We are able to obtain 89.7% accuracy on the testing split, which also clearly outperforms the baseline of 84.2%.

7.1 Adversarial examples

Since real-world samples tend to be diverse in nature, it is important to deploy models that are robust to distribution shifts in the data. One important shift are adversarial samples, which are samples that are imperceptibly different from a given image to humans, but since it uses the gradients from the models, to try to trick it. A very simple way to generate adversarial samples is using the Fast Gradient Sign Method (FGSM) [14]. A short PyTorch code-snippet on how to generate such as sample is shown below:

```
imgs = torch.sign(imgs.grad.data) * (3e-3) + imgs
```

Where 3e-3 is a common hyperparameter chosen since 3e-3 ~= 1/255 which is the value of one pixel. The adversarial accuracy of the model is 74.1% on the adversarially generated test set.

7.2 Robustness to Random Gaussian Noise

We also test the model on its ability to generalize and be robust to random Gaussian noise. Random Gaussian noise can be added to the image by:

```
imgs += torch.randn_like(imgs) * 3e-3
```

Such additive noise can also be added while taking or storing the image. The model accuracy to the random Gaussian noise is 88.6%. Random Gaussian noise is a more realistic test for the model, since adversarial samples need access to model parameters, and therefore is a "worst-case analysis".

8. Qualitative Results

To qualitatively examine the results, we used GradCam++ on the model that performed the best on the validation set [15]. The results are in figure 7, where the original image and the saliency map can be seen. The first and the second columns are samples from the unseen test set of a benign and malignant sample, respectively. We can clearly see that the model is *attending* to the correct parts of the image as the skin lesions are the epicenter of the focus of the model, and other spurious features such as hairs are what the model is attending to. The sections of the image that the model is paying the most attention to are the sample sections that a doctor may also closely examine, which is sensible. The model was also correct on determining the label for each of the two samples presented.

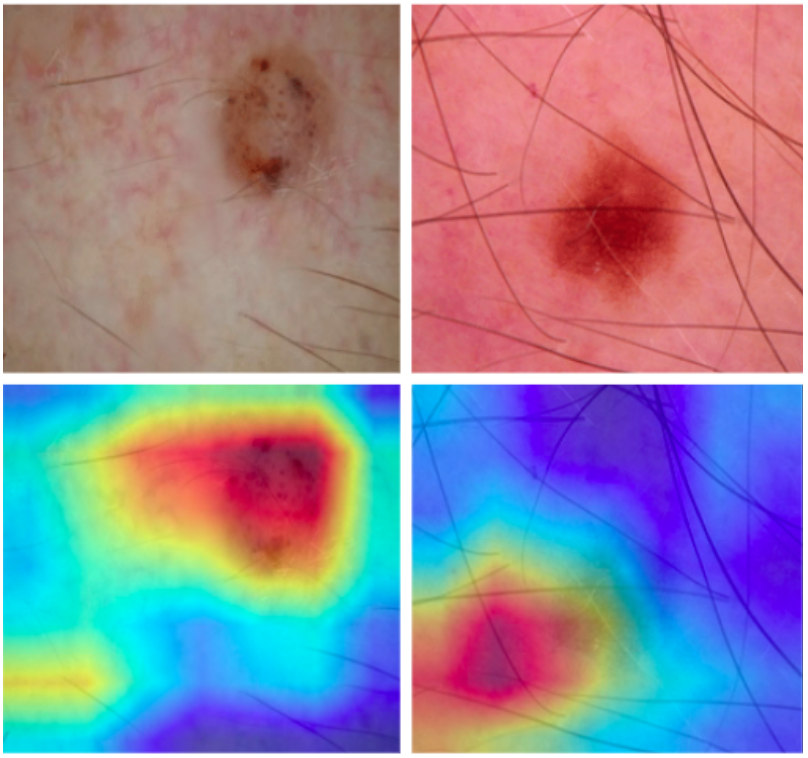

Figure 7: (left) A sample of a benign skin lesion and the corresponding attention map. (right) A sample of a malignant skin lesion and the corresponding attention map.

9. Evaluation on New Data

As our dataset was composed solely of skin lesions on lighter skin tones, it was primordial for our model to be evaluated on a sample dataset of darker skin tones. Skin cancer can affect anybody, thus we needed to accurately measure whether our model was biased towards lighter skin. An unseen dataset of malignant skin lesions on darker skin was created manually by searching through various dermatological journals [7][8][9][10]. From testing our baseline model on this new dataset, we concluded that our baseline model was biased towards lighter skin as it performed poorly on our manually created darker skin dataset. Thus, we did not expect our final model to perform much better since it was trained on the same dataset as our baseline model.

In comparison to our baseline model, our ResNet18 model performed significantly better on this unseen dataset, correctly predicting 8/9 labels. We decided to use saliency/attention maps in order to see why our model had performed so well on this dataset, given that our baseline had performed so poorly. In figure 7, it can be seen that the ResNet18 model correctly dissected the features of the skin lesions on lighter skin but from figure 8 it can be seen that

the model incorrectly focused on the upper right hand corner of the image.

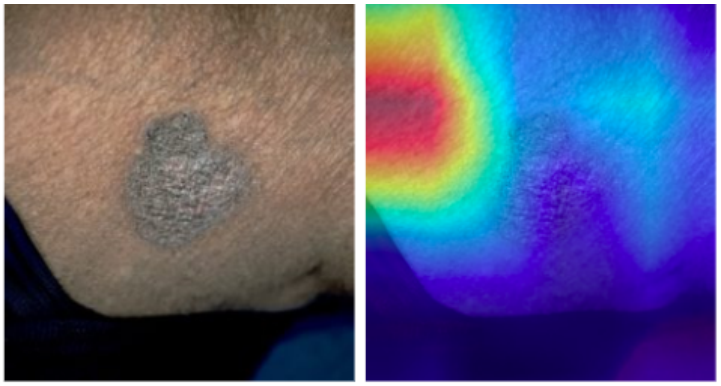

Figure 8. A sample of an unseen image with a significantly darker skin than the light-skin sample that the classifiers were trained on

Hence, it can be concluded that our model most likely achieved this 8/9 accuracy by simply guessing the labels rather than by actually extrapolating the features of the skin lesions on darker skin.

10. Discussion

As mentioned before, the proposed ResNet 18 model is able to outperform the baseline, and 6 other popular neural network architecture benchmarks. We use the model with the highest validation accuracy, and train it on the test set, and obtain a final test accuracy of 89.7%. Strategies such as L2 regularization or label smoothing could have been used to increase the ability of the models to generalize to the unseen test set. Furthermore, it's possible that the different architectures or different non-ImageNet pretrained variants could have further improved the performance of the models. Specifically, a model that is pre-trained on a biomedical dataset, such as a large-scale skin dataset, could have also been used to improve the model further.

We also noticed that the learned models are robust to random Gaussian noise added to the images, but significantly less robust to adversarial samples. This is also another important finding since for safety-critical applications, models must be robust to such spurious correlations as their ability to generalize is important for diagnostic purposes. Furthermore, using quantitative and qualitative results, we observed that although the proposed model is able to correctly classify dark-skin samples, after only being trained on light-skin samples. This may however be misleading, since the GradCam shows that the model is *attending* to random parts of the image, and not the skin lesion itself. Such consideration is important during the data gathering and training stages, on ways to make the dataset more diverse or the model less sensitive to overfitting. Finally, we also noticed in Table 1 on how shallower models with fewer parameters tend to outperform their deeper variants and do not generalize as well to the validation split. We hypothesize that models that are too large overfit the data which hinders their ability to generalize to unseen data. It is possible if we were to use a larger dataset to finetune the models the larger datasets would not overfit and perform better.

11. Ethical Considerations

Since our database provides us with images at a dermatoscopic level, the ethnicity of the patient could potentially affect the performance of our model. Because our database in particular includes mainly pictures of skin lesions on lighter skin, we have performed testing to investigate how the model might be biased against darker skin tones. Our results showed that there is currently a high degree of bias against darker skin and the model is guessing labels. This problem could hypothetically be solved by adding an equal amount of dark skin tone samples to our training dataset; however this is currently difficult to do because skin cancer is much rarer on darker skin tones and there is not enough data available at the moment [16].

12. Link to Google Drive

https://drive.google.com/drive/folders/1kQc0pxs155JfsnmkxIYOK3gbjUX-rfkk?usp=sharing

13. References

[1] "Skin cancer statistics", World Cancer Research Fund International. 2018 [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/skin-cancer-statistics

[2] "How is Skin Cancer Diagnosed?", SkinCancer.net, 2017. [Online]. Available: https://skincancer.net/diagnosis

[3] "Dermatologist-level classification of skin cancer with deep neural networks", Esteva, A., Kuprel, B., Novoa, R. et al., Nature 542, 115–118. 2017 [Online]. Available: https://doi.org/10.1038/nature21056

[4] N. Gouda and A. Vishwa Vidyapeetham, "Skin Cancer Classification Using ResNet", *Imedpub.com*, 2020. [Online]. Available: https://www.imedpub.com/articles/skin-cancer-classification-using-resnet.php?aid=28628

[5] "A patient-centric dataset of images and metadata for identifying melanomas using clinical context.", Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J. & Soyer, P., Sci Data 8, 34. 2021 [Online] Available: https://doi.org/10.1038/s41597-021-00815-z

[6] C. Fanconi, "Skin Cancer: Malignant vs. Benign," Kaggle, 19-Jun-2019. [Online]. Available: https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign

[7] H. Gloster and K. Neal, "Skin cancer in skin of color", *Journal of the American Academy of Dermatology*, vol. 55, no. 5, pp. 741-760, 2006.

[8] O. Agbai, K. Buster, M. Sanchez, C. Hernandez, R. Kundu, M. Chiu, W. Roberts, Z. Draelos, R. Bhushan, S. Taylor and H. Lim, "Skin cancer and photoprotection in people of color: A review and recommendations for physicians and the public", *Journal of the American Academy of Dermatology*, vol. 70, no. 4, pp. 748-762, 2014.

[9] "Skin Cancer & Skin of Color - The Skin Cancer Foundation", *The Skin Cancer Foundation*, 2021. [Online]. Available: https://www.skincancer.org/skin-cancer-information/skin-cancer-skin-of-color/

[10] "Skin Cancer Pictures & Photos | Pictures of Skin Cancer", *Cancer.org*, 2021. [Online]. Available: https://www.cancer.org/cancer/skin-cancer/skin-cancer-image-gallery.html?filter=Basal%20Cell%20Carcinoma,Kaposi%20Sarcoma,Melanoma,Merkel%20Cell%20Carcinoma,Skin%20Lymphoma,Squamous%20Cell%20Carcinoma

[11] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *arXiv.org*, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated Residual Transformations for Deep Neural Networks", *arXiv.org*, 2016. [Online]. Available: https://arxiv.org/abs/1611.05431

[13] G. Huang, Z. Liu, L. van der Maaten and K. Weinberger, "Densely Connected Convolutional Networks", *arXiv.org*, 2016. [Online]. Available: https://arxiv.org/abs/1608.06993

[14] I. Goodfellow, J. Shlens and C. Szegedy, "Explaining and Harnessing Adversarial Examples", *arXiv.org*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6572

[15] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", 2016 [Online]. Available: https://arxiv.org/abs/1610.02391

[16] M. Brennard and K. Buster, "Melanoma | Skin Cancer in Skin of Color | Melanoma in blacks", *Skin of Color Society*. [Online]. Available: https://skinofcolorsociety.org/dermatology-education/melanoma/#:~:text=Darker%20skinned%20people%20all%20have,have%20higher%20mortality%20than%20whites