DEEP LEARNING TECHNIQUES FOR MELANOMA CLASSIFICATION

Minnie Liang
West Lafayette High School
MehtA+Tutoring

Preeti Gomathinayagam
Dougherty Valley High School
MehtA+Tutoring

Mohammad Sharafat
Al Ittihad Pvt School
MehtA+Tutoring

Isabelle Hu
Middlesex School
MehtA+Tutoring

May 27, 2024

ABSTRACT

The most common cancer globally is skin cancer. Early detection of skin cancer can drastically increase patient survival rates; therefore, a computerized image classification system of skin lesions can save time, and by extension, human life. In this paper, we elevate a traditional CNN model which inputs only images into a state-of-the-art multimodal model which concatenates the CNN image model with metadata features. Our results show that our multimodal model outperforms the unimodal model by a 12.15% increase in accuracy on average. We further improve our model by exploring various CNN architectures, specifically ResNet-18 and VGG16. Our accuracies increased by 9.81% on average when using ResNet-18, and we confirm these results by applying the Grad-CAM algorithm on our skin lesion images.

Keywords: Skin cancer; Multimodal deep learning; Image augmentation; Convolutional neural network; Saliency maps; Grad-CAM

1 Introduction

Skin cancer is the most common cancer worldwide. In the United States, 1 in 5 Americans will develop skin cancer by the age of 70, and more than 2 Americans die of skin cancer every hour [1]. Occurrences of melanoma, the deadliest form of skin cancer, are increasing faster than any other preventable cancer in the United States [2]. Melanoma occurs when melanocytes, which produce the melanin that gives skin a tan or brown tint, start growing rapidly out of control due to mutations in genes controlling cell growth [3]. Although less common than other skin cancers, melanoma is the most aggressive skin cancer due to its rapid growth rate [4]. It is highly probable that melanoma cancer cells will grow and spread to other parts of the body if not detected at an early stage and treated. However, melanoma has a 99% 5-year survival rate when detected early [1]. Early detection is crucial to providing melanoma patients with the highest possibility of survival without relapse.

However, recent studies have shown that there are approximately 3.3 dermatologists for every 100,000 people [5], which means that there are simply not enough practicing dermatologists available to properly serve all communities. With the dermatologist shortage issue, another problem arises in the unreasonably long wait times before even seeing a dermatologist. On average, there is a 38 day wait time between booking the appointment to actually seeing the dermatologist, and the wait time further increases in more populous cities [6]. Long wait times for dermatology appointments negatively affect both the patient's experience and the patient's safety [7].

Therefore, we propose that deep learning classification of melanoma in its early stages may help reduce patient wait times by filtering out patients who do not require a dermatologist's assistance. Patients who do have lesions of concern (classified as malignant by the model) can then be referred to a dermatologist in an efficient manner, allowing for earlier detection and higher quality of care. In this paper, multiple deep learning techniques for melanoma classification are proposed, tested, and implemented using a publicly available skin lesion image dataset.

2 Related Work

Many of the current methods of implementing machine learning for melanoma classification focus on a CNN model that solely uses images to classify skin lesions. There have been various advances in using images to predict melanoma, such as applying a three-level fusion approach of the CNN models DenseNet-121, ResNet-18, and ResNet-50 [8], implementing a squeeze-and-excitation network, and combining semi-supervised learning with Mean Teacher Method [9]. However, these methods present large drawbacks: by only using images, they neglect crucial pieces of information such as age, gender, and anatomical site that are used by dermatologists when diagnosing. With this paper, we go beyond the image-only traditional CNN. Instead, we concatenate images with metadata side by side, creating a more precise, state-of-the-art multimodal model. Additionally, there has been recent research on the accuracies of various data augmentation techniques, like between-class learning, random erasing data augmentation, body hair augmentation [9], and using GANS to generate images of varying styles [10]. With these recent data augmentation advances, we also implement them in our model to augment our dataset and achieve higher accuracies on our model.

3 Background

3.1 Dataset

There are few skin lesion image datasets available to the public which are of adequate size for a machine learning algorithm to train on. There are even fewer datasets specific to melanoma. We considered multiple different datasets, but ultimately did not choose them, such as:

- Dermofit Image Library [11] is a dataset containing 13,000 images of 10 different classes. 76 of the images in this dataset are of melanoma; not nearly enough for our model to train on.
- Dermnet [12] is a skin disease atlas with images of a wide variety of skin conditions. However, the images of melanoma are watermarked, so the group chose not to implement this dataset into the model.
- HAM10000 [13] is a dataset with a total of 10,015 images of various diagnostic categories such as basal cell carinoma, squamous cell carcinoma, and melanoma. 1,113 images from this dataset are of melanoma. We did not choose this dataset as it was a subsection of the dataset we chose.

The dataset used was the ISIC 2020 Challenge Dataset generated by the International Skin Imaging Collaboration (ISIC) using images from Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, The University of Queensland, and the University of Athens Medical School. The dataset consists of skin lesion images, patient ID, gender, age, anatomical site, and lesion diagnosis. Each image is classified as either benign or malignant melanoma. Examples of various images in the dataset are included in Figure 1.

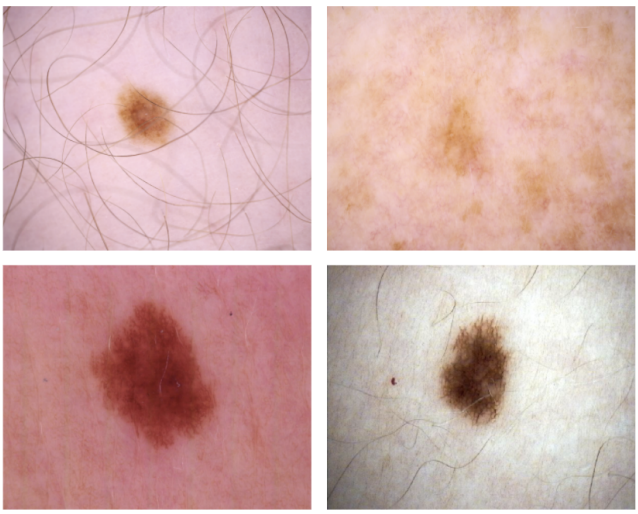


Figure 1: Examples of skin lesion images in the dataset

4 Methodology

4.1 Metadata Preprocessing

In order to correct gaps in the data and convert it into an interpretable format, we removed rows with missing values in the gender column and convert the gender values into binary (0 for female, 1 for male). To correct the age variances in the metadata, we normalized the age to a [0,1] range. Age is normalized by dividing all the ages by the maximum age, 90. In addition, the anatomical site of the skin lesion is encoded into one hot vectors of either 0 (not the lesion site) or 1 (lesion site). There were six possible anatomical sites: head/neck, lower extremity, oral/genital, palms/soles, torso, and upper extremity, or unknown. In Figure 2, we can see how skin lesions are indeed impacted by the anatomical site [14]. The majority of lesions come from the torso, and this intuitively makes sense as the torso is frequently exposed to the sun. Additionally, only a few percentage of the skin lesions come from the palms and soles, which is logical as these areas rarely make contact with the sun. From here, we created a feature vector of size nine with the characteristics of gender, age, and anatomical site. For example, the feature vector of [1, 0.28, 0, 0, 0, 0, 1, 0, 0] represents an approximate 25-year-old male with a skin lesion on the torso.

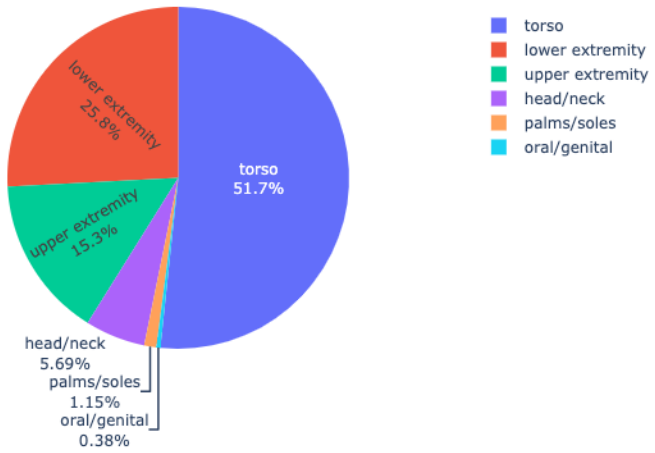


Figure 2: Pie Chart of the Anatomical Site

There is a massive data imbalance between the number of benign and malignant images. The original dataset contains 32,542 benign images and 584 malignant images, so approximately 1.76% of the data is malignant. To combat the data imbalance, we took a sample of size 1850 from the benign images. We then split this data into a train set and a validation set, with 463 benign and 146 malignant images in the validation set. For the train set, we started out with 1387 benign images and 438 malignant photos. However, by performing hair image augmentation twice, we were able to triple the number of malignant images in the train set, bringing us to an approximately even number of 1,387 benign and 1,314 malignant images. While augmenting each malignant image, we needed to ensure that the feature vector for the image was also kept with the newly augmented image. We handled this by adding a separate column in the feature vector CSV file to distinguish augmented from non-augmented images (0 represented non-augmented images and 1 or 2 represented augmented). As shown in Table 1, our train set included missing values for anatomical site, age, and sex. However, the test set contained missing values for only the anatomical site. Since the test set did not include any missing values for age or sex, we decided to remove rows from the train set that contained missing values for age and sex. It was unnecessary to remove any rows missing only values for anatomical site because the test set also contained missing values for anatomical site, so there would not be a significant impact.

Table 1: NaNs of Dataset

	Feature	Missing Values	% of total values
Training	Anatomical Site	527	1.6
	Age	66	0.2
	Sex	65	0.2
Validation	Anatomical Site	351	3.2

4.2 Image Preprocessing

Input images are preprocessed by: (I) resizing/cropping the image to 224x224 pixels; (II) converting images to [C,H,W] tensors; (III) normalizing tensors to values with a given mean/standard deviation. Figure 3 shows examples of images resized to 224x224 pixels and the impact of normalizing the tensors on the images. We avoid converting our images into greyscale as the conversion limits the model's ability to learn from the image, and this is because color is a key factor in diagnosing melanoma.[15]

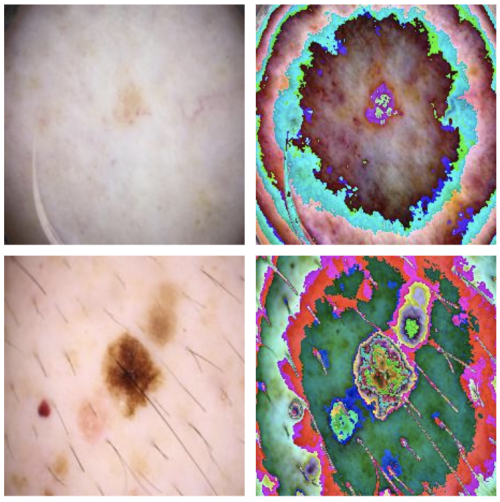


Figure 3: Resizing and normalizing the images

4.3 Image Augmentation

The malignant images are augmented to increase image diversity and to decrease the data imbalance between benign and malignant images. However, when dealing with medical images, the safety of image augmentation must be taken into account [10]. For example, using methods that change the color pigmentation of the images will endanger the images' target label; an image previously classified as melanoma transformed by color pigmentation might give out a false positive. Therefore, the melanoma images are augmented using a safe form of image augmentation, and augmentation methods such as transformations are avoided to avoid risking the label's integrity [16].

- Hair Augmentation: The melanoma images are augmented by concatenating melanoma images with predefined hair strand images (Figure 4). The size or width of each hair strand is randomly resized. The number of hairs being added to each augmented image is randomly chosen, and the position of the hairs onto the skin lesion are randomly placed. This approach increases the diversity of the dataset while being safe, thus resulting in higher accuracy and reduced overfitting. [17]

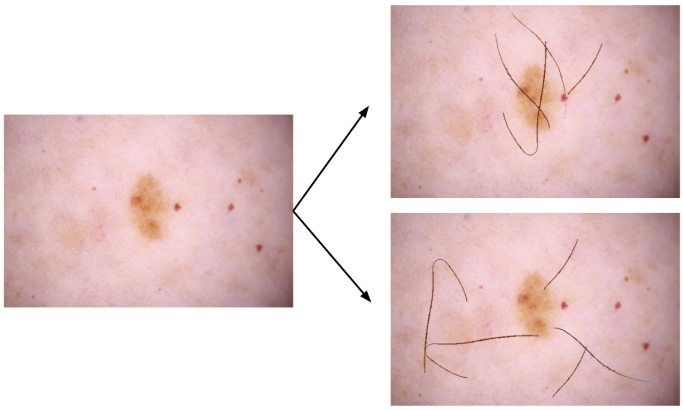


Figure 4: Malignant Image Hair Augmentation

4.4 Model

We implemented two models, one where we predicted melanoma classification based only on the images [18], and a second where we created a multimodal model by combining the images and their respective metadata. A structure of the model used when only images were inputted is shown in Figure 5. We took a CNN, first ResNet-18, then VGG16, and constructed classifier architecture on top. The classifying infrastructure consisted of three fully connected layers and had an activation function of LogSoftMax which resulted in two outputs.[19]

In addition to the unimodal model with only images, we also constructed a multimodal model consisting of images and additional features [20]. We concatenated the CNN model, first ResNet-18, then VGG16, with the metadata features side by side. The classifying infrastructure included 3 fully connected layers and a sigmoid activation function, resulting in one output, either benign or malignant. A diagram of this model's architecture is shown in Figure 6.

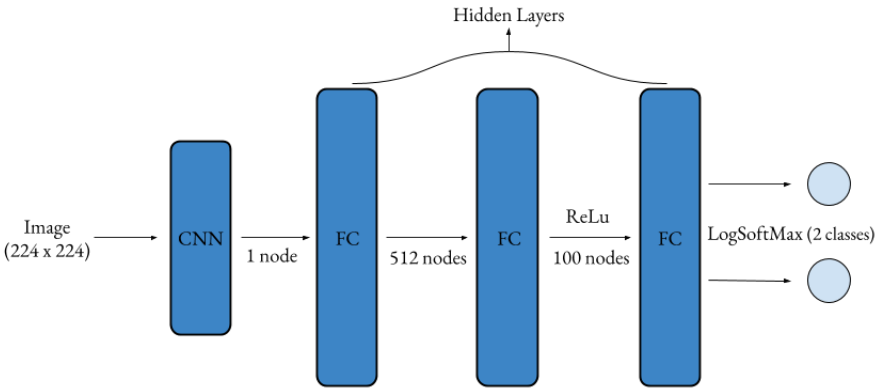


Figure 5: CNN only

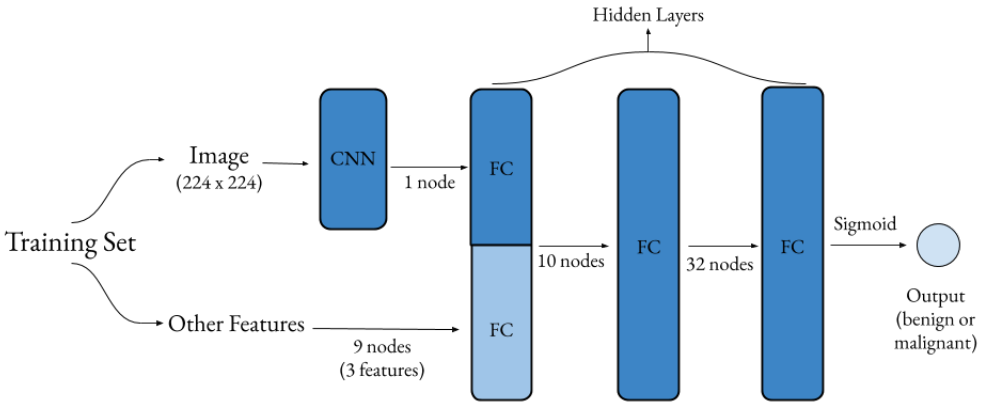


Figure 6: CNN with additional features

The loss function used for the images only CNN model was Cross Entropy Loss, and is given by

$$L = -(y \log(p) + (1 - y) \log(1 - p)). \tag{1}$$

Our loss function for the multimodal model was a Binary Cross Entropy with logits loss, which combines the sigmoid layer and the Binary Cross Entropy loss. This method of combining the sigmoid layer and the BCE loss into one layer is much more numerically stable than separating the sigmoid and BCE loss into separate layers. The loss function was defined as

$$\ell(x, y) = L = \{l_1, l_2, ..., l_N\}^\top, \tag{2}$$

$$l_n = -w_n[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))] \tag{3}$$

where N is the batch size. Summing up the logs of these exponentials can be advantageously used for this loss function to provide numerical stability.

4.5 Experiments

We compared two different CNN architectures, ResNet-18 and VGG16, for both the unimodal and multimodal model. To ensure other parameters were not affecting our accuracies, we kept the batch size of 32 and learning rate of 0.01 constant throughout all the models. We also maintained the same optimizer, Adam, for the unimodal model, and the same optimizer, Stochastic Gradient Descent (SGD), for the multimodal model on all of our runs.

5 Results

5.1 VGG16

From all the epochs ran on the unimodal VGG16 model, the highest train and validation accuracy observed was 73% and 77%, respectively. From Figure 7, we see that in the beginning, the model's validation accuracy oscillated slightly, before evening out as more epochs went on. The training and validation accuracies stabilize to approximately 70% and 60% respectively. We chose the model at epoch 7, which had a 70% training accuracy and 64% validation accuracy. The resulting test accuracy was 62.86%.

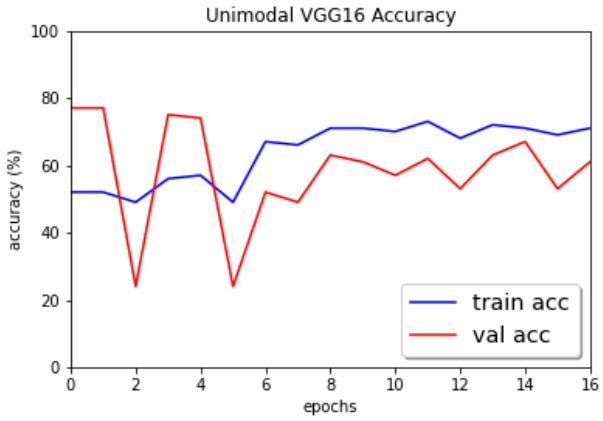


Figure 7

The confusion matrices shown in Figure 8 are constructed for the training and validation datasets. Even though the test accuracy is not high, we see that the most common error made by our model was classifying benign lesions as malignant; this false positive is the ideal error as it is better to be safe than sorry when dealing with medical diagnoses of a potentially life-threatening condition. However, this model is not optimal as it classifies a large proportion of benign images as malignant.

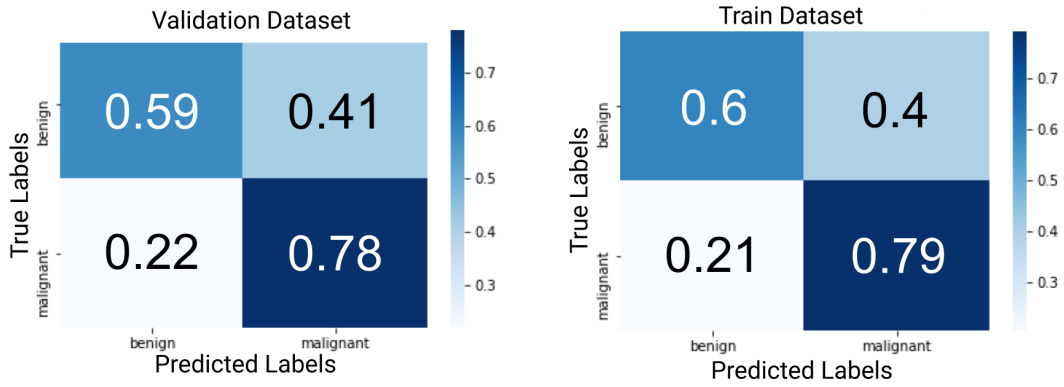


Figure 8: Confusion Matrices for Unimodal VGG16 on Epoch 7

5.2 ResNet-18

For all the epochs we ran on the unimodal ResNet-18 model, the highest train accuracy achieved was 77% and the highest validation accuracy achieved was 70%. We chose the model at epoch 27, which had a 77% train accuracy and 69% validation accuracy. Our model yielded a test accuracy of 70.65%. We can see in Figure 9 that the unimodel ResNet-18 does not deviate much from around a 75% train accuracy and a 66% validation accuracy, although there are a few areas of the graph where the model is likely entering a local maximum for loss.

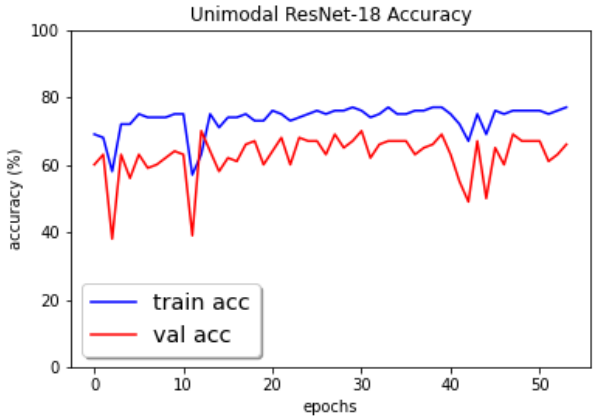


Figure 9

Confusion matrices constructed on the training and validation dataset for the ResNet-18 model at epoch 27 are shown in Figure 10. In both confusion matrices, we see that our model produces more false positives than false negatives: the ideal error.

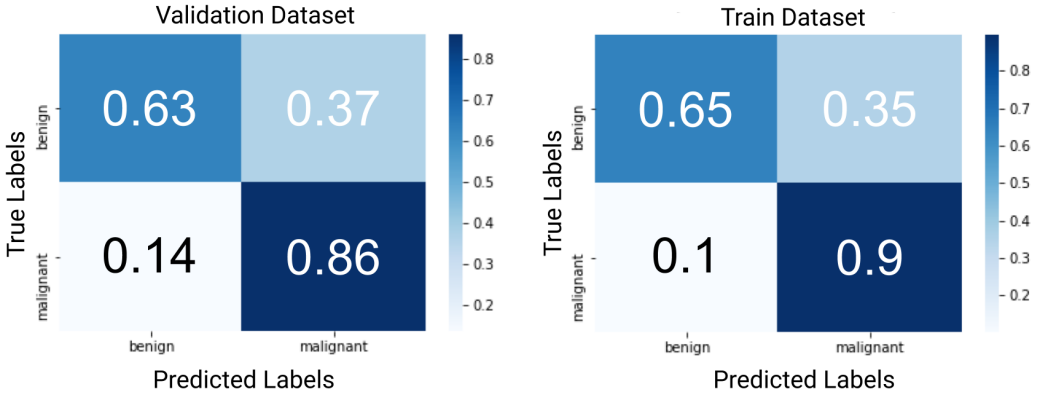


Figure 10: Confusion Matrices for Unimodal ResNet-18 on Epoch 27

5.3 Multimodal VGG16

Out of all the epochs ran on the multimodal VGG16 model, the highest train and the highest validation accuracy obtained were 94.85% and 77.83%, respectively. Figure 11 shows that as the model trains on more and more epochs, the validation accuracy levels off to about 75.1%. We chose two different epochs for this model: epoch 12 and epoch 24. The model at epoch 12 had a 81.52% train accuracy and a 76.52% validation accuracy, producing a 72.04% test accuracy. In comparison, the model at epoch 24 had a 91.07% train accuracy and a 77.83% validation accuracy, resulting in a 72.15% test accuracy.

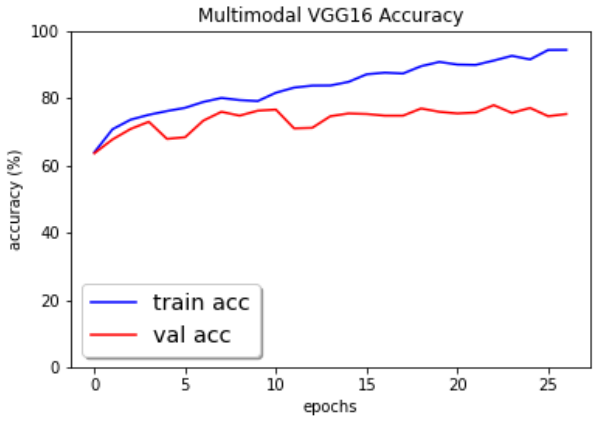


Figure 11

Taking a look at Figure 12, we can see that the VGG16 multimodal model at epoch 24 is not the best model from a real world standpoint. The model may have a relatively high validation accuracy of 77.83%, but it is consistently miscalculating when classifying a malignant skin lesion. The false negative rate is extremely high, which could be fatal as we are dealing with such an aggressive disease where early detection is crucial.

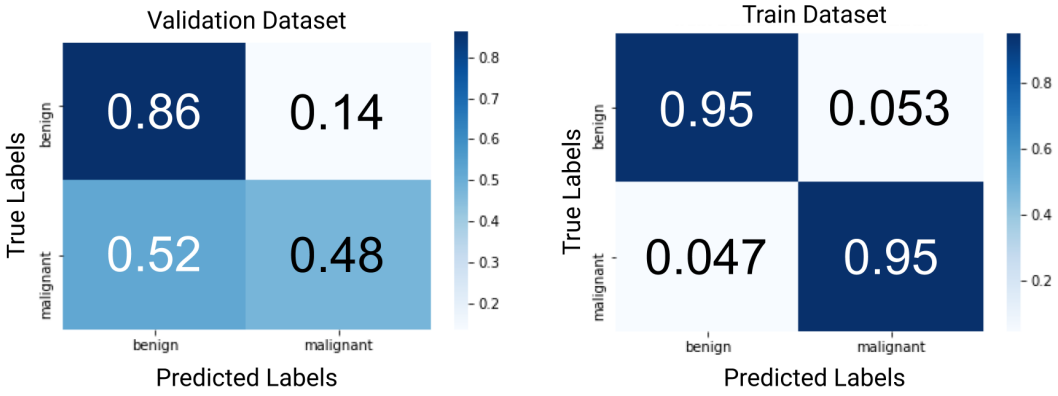


Figure 12: Confusion Matrices for Multimodal VGG16 on Epoch 24

However, if we compare epoch 24 with epoch 12, despite the small difference between their testing accuracies - which is 0.11% - we can see in Figure 13 that the multimodal VGG16 model at epoch 12 a better model for real world application. The training accuracy may be much lower than the training accuracy of epoch 24, but the model at epoch 12 predicts the classification of a malignant lesion with higher accuracy than the model at epoch 24. From epoch 24 to epoch 12, the false negative rate also decreases significantly, helping to increase early detection of melanoma and thus potentially increasing the patient survival rate of melanoma.

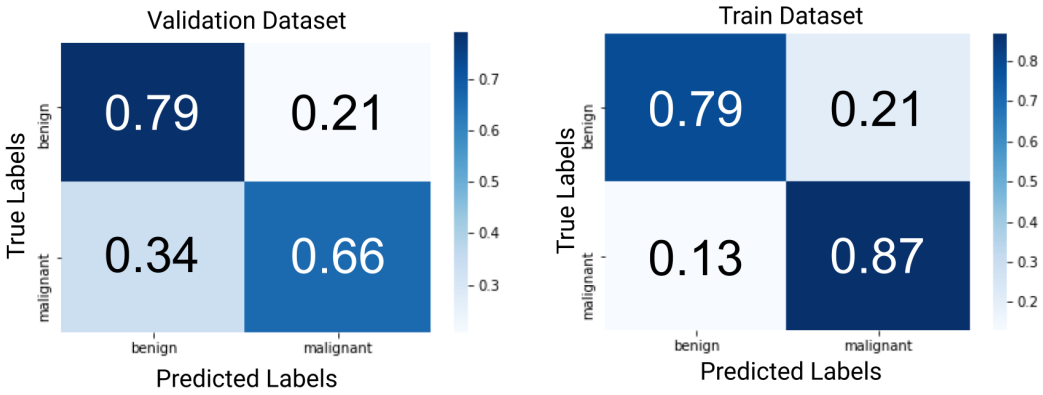


Figure 13: Confusion Matrices for Multimodal VGG16 on Epoch 12

5.4 Multimodal ResNet-18

For all the epochs ran on the multimodal ResNet-18 model, the highest train accuracy achieved was 99.7408% and the highest validation accuracy achieved was 80.1314%. We chose the model at epoch 13, which had a 87.63% train accuracy and a 77.6683% validation accuracy. In Figure 14, we see that the training accuracy converges to almost 100% and the validation accuracy converges to around 77.3%.

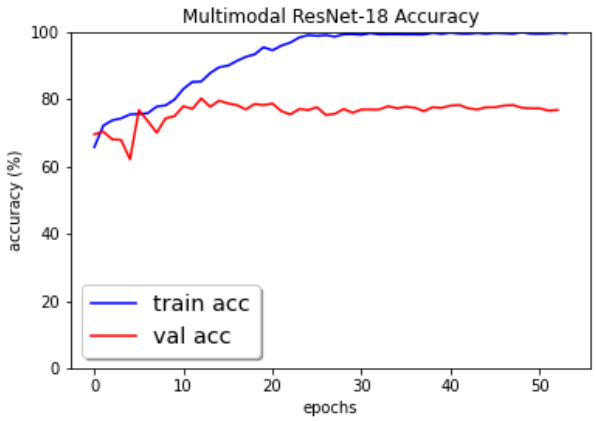


Figure 14

While these high train accuracies at higher epochs seem encouraging, if we take a look at the confusion matrices for epoch 23 in Figure 15, we see that these higher epochs are not ideal for a real-life applicable model. The majority of the errors happening at these higher epochs are at false negatives, which is extremely harmful for patients as it can compromise early detection of melanoma.

In contrast, the model at epoch 13 in Figure 16 shows that the multimodal ResNet-18 model does not overly tend to make a false negative error. Instead, the model slightly tends to make a false positive error, which again, is the safer and more desirable error. At epoch 13, our model produced a 77.36% test accuracy, whereas the model at epoch 23 only produced a 74.99% test accuracy.

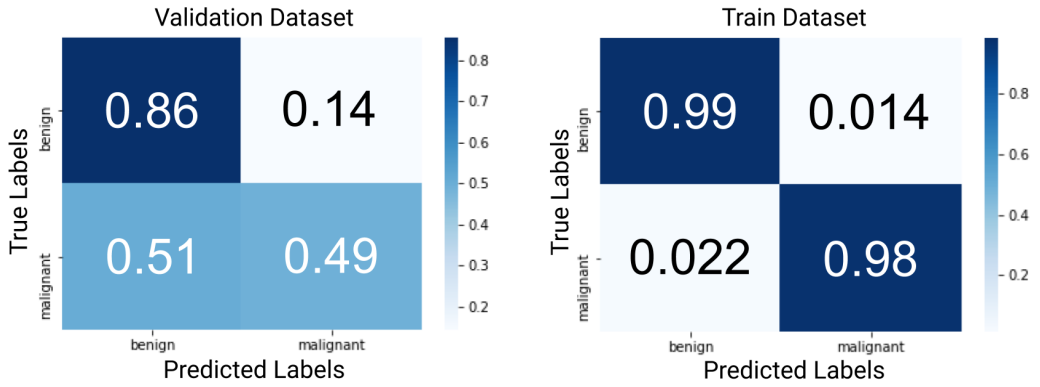


Figure 15: Confusion Matrices for Multimodal ResNet-18 on Epoch 23

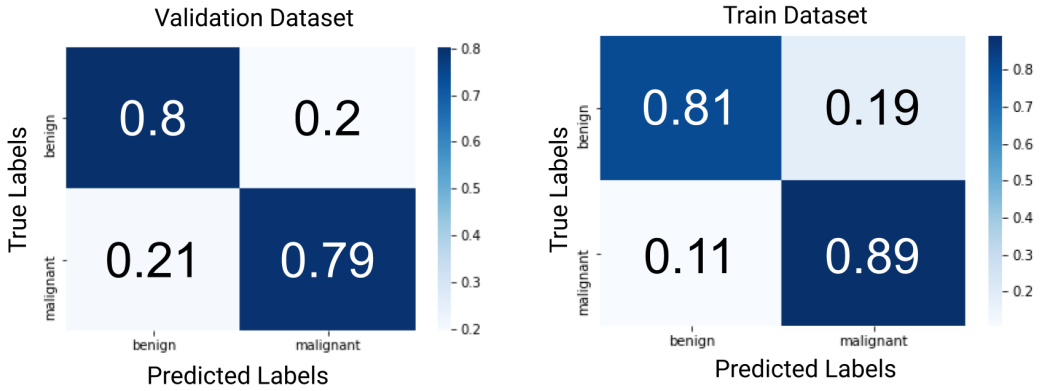


Figure 16: Confusion Matrices for Multimodal ResNet-18 on Epoch 13

By using ResNet-18 instead of VGG16, our multimodal model saw a 7.22% increase in accuracy. The test accuracy for the unimodal model increased by 12.39% when using ResNet-18 compared to VGG-16. In addition, ResNet-18 is also significantly less computationally expensive to run than VGG16. This stems from the idea that overall, the speed of a convolution depends on the size of the input image. In our case, the layers of ResNet-18's architecture reduces the height and width of the input image at a much faster rate than VGG16's architecture.

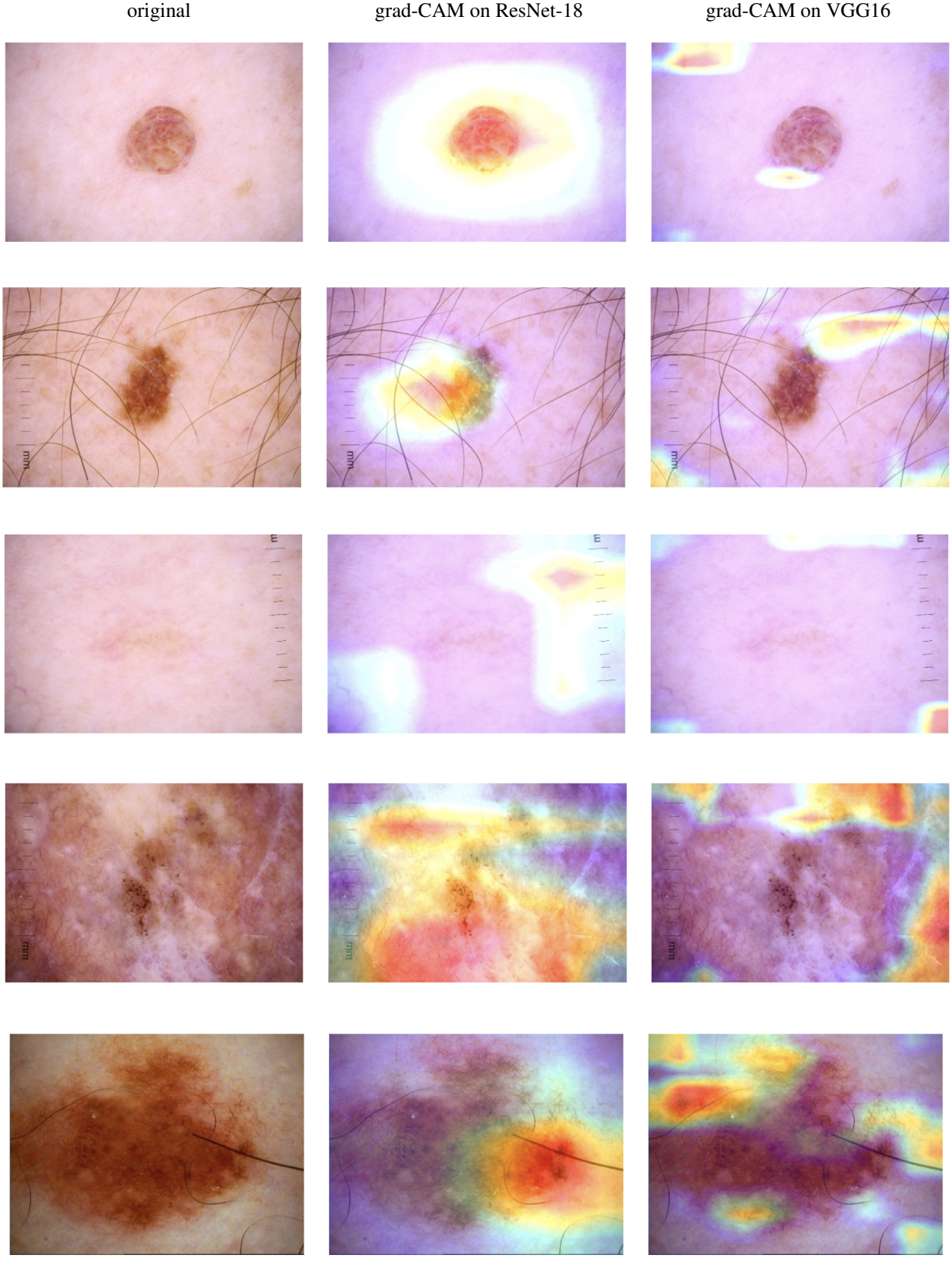
Our state-of-the-art multimodal model exceeded in increasing the accuracy of malignancy skin lesion predictions compared to the traditional unimodal CNN model. There was a 9.5% increase in accuracy by switching over from the unimodal ResNet-18 to multimodal ResNet-18, and a 14.8% increase in accuracy by switching over from unimodal VGG16 to multimodal VGG16.

6 Interpretability

6.1 Grad-CAM

We implemented the Grad-CAM algorithm and saliency maps on the skin images to determine which areas of the image were influencing the model's decision in benign/malignant classification. Grad-CAM is a method of using the model's final convolutional layer's gradients to generate a course localization map that highlights the important regions in the image for prediction [21, 22]. The grad-CAM algorithm confirmed our result that ResNet-18 yields a higher accuracy than VGG16. Figure 17 shows that ResNet-18 uses the area surrounding the skin lesion more often to determine a classification than VGG16. However, the third row of images shows our models' imperfections as they have difficulties detecting the lesion when it is faint and not as apparent. Note that the model has a harder time when the lesion is more widespread than when the lesion is smaller and concentrated in one area when comparing the first and second row with the fourth and fifth row.

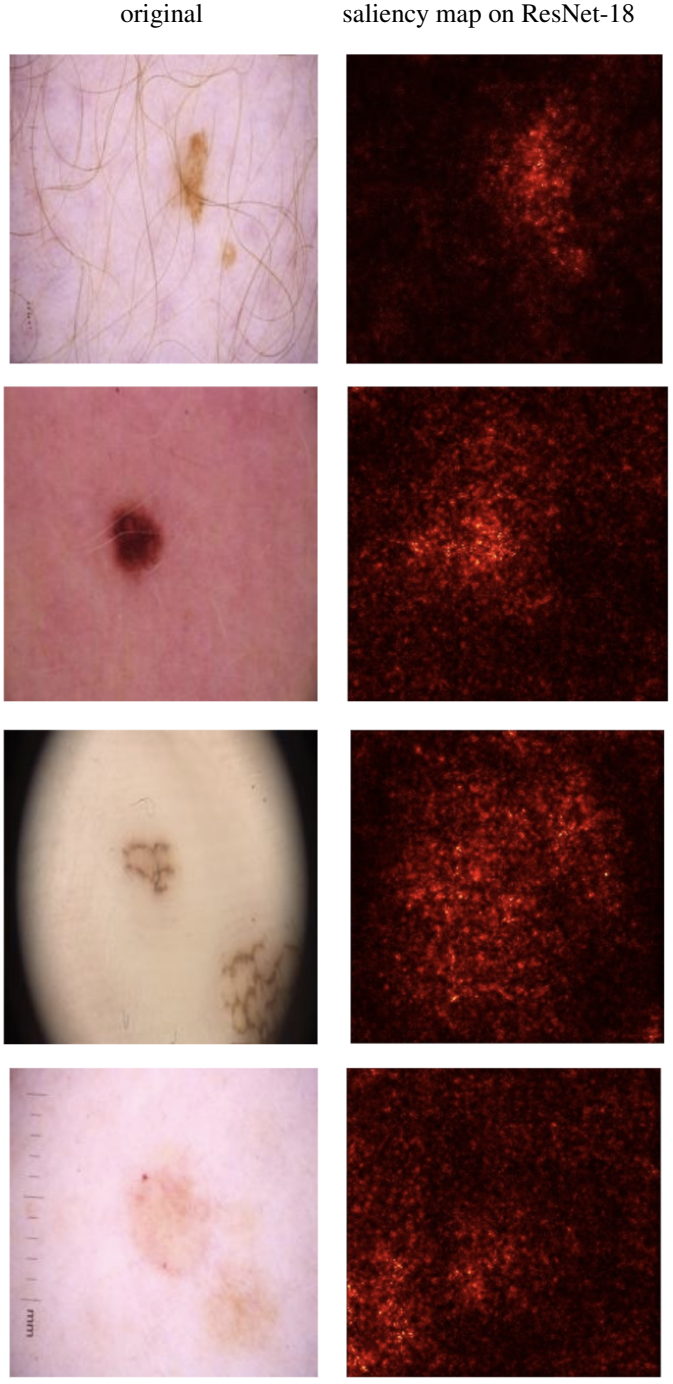
Figure 17



6.2 Saliency Maps

In addition, we implemented saliency maps: heatmaps that highlight the areas of the image that contribute the most to the predicted classification [23, 24]. Saliency maps are similar to Grad-CAMS, and give insight to the "black box" issue of machine learning. Looking at Figure 18, we observe similar patterns in how our model chooses the most influential areas to contribute to classification. Small lesions focused in one area are easier for the model to detect, whereas fainter lesions which might blend in more with the background skin are more difficult to detect.

Figure 18



7 Future Work

One of the limitations of the ISIC 2020 Challenge Dataset was its small number of melanoma images. While we overcame this limitation through hair augmentation, we still had very few malignant images in comparison to total benign images. Future work would involve developing additional image augmentation techniques while keeping concerns about maintaining the integrity of the image classifications in mind.

In addition, the majority of the datasets found lacked racial diversity. This may be due to the prevalence of melanoma in light skinned populations as people with more melanin are at decreased risk for skin cancer [25]. The pigmentation of melanoma in non-white populations differs enough that another model would need to be constructed on a different dataset in order to properly classify images as benign or malignant.

8 Conclusion

In this paper, we present a conventional unimodal model and compare it with an innovative multimodal model for classification of malignancy of skin lesions. The results of our study show that the multimodal model yields higher accuracies than the unimodal model. We also implement two different CNNs, ResNet-18 and VGG16, and conclude that ResNet-18 produces higher model accuracies. In addition, we investigate the effect of image augmentation techniques on the integrity of the image's label, and demonstrate that the Hair Augmentation Technique is the safest.

9 Division of Labor

We divided the work as follows:

- Inputting the Dataset - Minnie Liang and Isabelle Hu
- Preprocessing the Dataset - Preeti Gomathinayagam and Mohammad Masoud
- Modeling- Minnie Liang, Isabelle Hu, Preeti Gomathinayagam, Mohammad Masoud
- Performance Metric- Minnie Liang, Mohammad Masoud, Isabelle Hu
- Visualizations- Isabelle Hu, Preeti Gomathinayagam, Mohammad Masoud, Minnie Liang
- Saliency Maps/GradCAMs- Minnie Liang, Preeti Gomathinayagam

10 Acknowledgements

We would like to thank Bhagirath Mehta, Haripriya Mehta, Marwa AlAlawi, and Andrea Jaba for their help in teaching and advising us throughout this project.

References

[1] Skin Cancer Facts and Statistics. April 2020.

[2] Susan Swetter MD Alan C Geller RN MPH. Screening and early detection of melanoma in adults and adolescents. *UpToDate*, May 2020. edited by Joann G Elmore, MD, MPH; Hensin Tsao, MD, PhD; Jane Givens, MD.

[3] Melanoma. July 2020.

[4] Allen Halpern, Ashfaq A. Marghoob, and Ofer Reiter. Melanoma overview.

[5] There's a reason you have to wait so long for a dermatologist appointment, Aug 2018.

[6] Diane Mapes. The dermatologist won't see you now, Mar 2007.

[7] Sara Heath. Long wait times in dermatology harm patient experience, safety. May 2019.

[8] Amirreza Mahbod, Gerald Schaefer, Chunliang Wang, Rupert Ecker, Georg Dorffner, and Isabella Ellinger. Investigating and exploiting image resolution for transfer learning-based skin lesion classification, Jun 2020.

[9] Yuexiang Li and Linlin Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2):556, 2018.

[10] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[11] Dermofit image libary.

[12] Dermnet: Malignant melanoma photos.

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

[14] Ibtesama. Siim baseline keras(vgg16), Jun 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Dec 2015.

[16] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, Dec 2017.

[17] Roman. Melanoma. pytorch starter. efficientnet, Jul 2020.

[18] Adria Romero-Lopez, Xavier Giro-I-Nieto, Jack Burdick, and Oge Marques. Skin lesion classification from dermoscopic images using deep learning techniques. *Biomedical Engineering*, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, page 630–645, 2016.

[20] Felipe Moreno, Haripriya Mehta, and Daniel Lee. Multimodal classification of lung diseases, Jul 2020.

[21] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[22] Stepan Ulyanin. Implementing grad-cam in pytorch, Feb 2019.

[23] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6021–6029, 2020.

[24] Aditya Rastogi. Visualizing neural networks using saliency maps in pytorch, Apr 2020.

[25] World Health Organization. Ultraviolet(uv) radiation and skin cancer, Oct 2017.