Skin Lesion Detection and Classification

Abha Godse
Department of Biomedical Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
abhag@andrew.cmu.edu

Brandan Dunham
Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA 15232
brd86@pitt.edu

Nour Jedidi
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
njedidi@andrew.cmu.edu

Rachita Jain
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15232
rjain2@andrew.cmu.edu

Abstract

This paper presents a novel approach towards the skin lesion detection and classification problem to help computer-aided diagnosis of skin cancer. We developed an ensemble of two state-of-the-art CNN models- ResNet-50 and EfficientNet-b0 and trained them on the publicly available ISIC 2019 classification challenge data set. GANs were used to generate more data for all the classes. The performance of the network was evaluated on various metrics of balanced multi-class accuracy.

1 Problem Statement

Skin cancer is a deadly disease and its incidences continue to increase. Melanoma is the most fatal kind of skin cancer and the early detection of such skin lesions can prevent their growth and can therefore aid timely medication.

Visual detection by expert dermatologists have proved to be beneficial in the skin cancer diagnosis, but to train these detection systems takes a lot of money and effort. The tremendous improvements in the field of automatic image analysis have led to a great deal of research in detecting the pigmented skin lesions from standard photographs as opposed to Dermoscopy. This computer-aided detection of pigmented lesions and their further classification into different types of cancers can speed up the cumbersome process of diagnosis.

In this task of "Skin Lesion Detection and Classification", we aim to develop an approach towards detecting the pigmented lesions and classifying them into 8 known categories of skin cancers and an unknown category as proposed by the ISIC challenge 2019. The data-set provided has a great deal of imbalance between classes and also does not suffice for any state-of-the-art deep learning methods. Both these challenges have been addressed and some pre-processing steps have been proposed.

2 Literature Review

A great deal of research has come out of the ISIC challenge in the last few years and various methods have been proposed. CNNs and their variations have seen an extensive use in most of the top scores of the challenge. The most successful models have used the challenges posed by classes' imbalance and less data have been addressed by using GAN networks. Pollastri et al. [2019] used DCGAN and LAPGAN models to generate more data. The model involved 6 networks used together as an

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

ensemble, and used RGB, HSV, and L* color channels for all images. Overall, LAPGANs show an improvement of about 1 per cent accuracy relative to the baseline model.

The pre-processing done on the images have also shown significant improvements in the overall accuracy. In Bisla et al. [2019] authors utilized a combination of scaling and masking to remove hair from melanoma images in order to prevent the network from learning to overfit using them. The masking approach is based on the paper "Towards a computer-aided diagnosis system for pigmented skin lesions", which converts pixel colors into an LUV representation and removes hair based on meeting a certain color threshold. Additionally, the authors used DCGANs to create virtual patients for rare classes. Overall, these approaches improved the accuracy of the network versus the top networks from the 2017 ISIC challenge at high sensitivity (>89 per cent) thresholds and overall AUROC when using a traditional ResNet-50 network pretrained with Imagenet data. The 2017 challenge included 3 categories total, and the final testing for this paper was performed on 600 images from the ISIC test set. Another paper by Schmid-Saugeona et al. [2003] contains information on the hair removal methodology from the above paper, in additional to calculations related to symmetry and other important metrics for melanoma diagnosis.

Federico Pollastri [2019] trained multiple neural networks using only data from the 2019 ISIC challenge competition, placing second among those without using external, private data sources. Networks were based on one of three architectures: DenseNet, ResNet, and SeResNext (squeeze and excite + a modified resnet architectures called ResNext). Additionally, training is done on 8 CNNs using a leave one class out methodology in order to detect classes that appear outside of the given classes learned, to detect the missing class in the 2019 ISIC challenge. Steven Zhou [2019] performed the best among algorithms without using external private data from the ISIC 2019 challenge, using multiple networks pretrained on ImageNet data. Their network used a total of eight different networks, DenseNet121 (3 times), SeResNext50, SeResNext101, EfficientNet B2, EfficientNet B3, and EfficientNet B4 as their base classifiers, and used a sample of three or four of them for each ensemble. To detect the missing class, it found images that had low probabilities as their best class (score below a threshold of 0.35) to select as being unknown rather than their top class. The EfficientNet and SeResNext architecture outperformed the Densenet architecture in cross validation, but all networks were tested in at least one ensemble. The final submitted ensembles were either exclusively EfficientNets, or EfficientNets with one DenseNet.

3 Methods

Models used by some of the top scoring submissions in the ISIC competition included ResNet, DenseNet, Squeeze and Excite ResNext, and EfficientNet.

For our data pre-processing, we rotated and flipped the images with a 50% probability. We also applied color normalization. This approach was used by several submissions, such as those by Parreno et. al and Rusong et. al, in the ISIC competition, and subsequently by our model. One major difference in the pre-processing cycle of different algorithms was the way the images were scaled and cropped to be used in the networks. For our algorithm, we tested and saw that scaling to 768 pixels and then cropping to 512 pixels outperformed scaling to 512 or 384 pixels and then cropping to 256 pixels. However, most of our models used scaling to 512 pixels on the smallest edge, followed by a random crop for training, or a center crop for validation of 256 x 256 pixels to allow us to run more experiments.

While competitors chose between several different sizes of these network types, such as EfficientNet-b2 or EfficientNet-b3, and DenseNet-121 or DenseNet-201, we focused primarily on EfficientNet-b0 – which had the best results in preliminary testing – and ResNet-50, which is arguably the most popular network structure for image classification used today. Most testing focused on EfficientNet-b0 due to it having the highest scores in preliminary research. To improve our initial results, we used pre-trained weights from ImageNet. For all classification networks, we used the Adam optimizer with a default learning rate of 1e-3. The learning rate was decreased by 10% in each epoch, for a maximum of 15 epochs run per test. All algorithms were trained with a batch-size of 16, using a weighted cross-entropy loss function to account for class imbalance.

Typically, with unbalanced classes, Deep Learning systems tend to struggle to learn as most models will tend to lean towards the majority class Bisla et al. [2019]. Additionally, the deeper the CNN, the more prone to the threat of vanishing gradients it is. Moreover, as the dataset is not adequate for

Table 1: Training Data by Class

Class	Image Count
Actinic Keratosis	867
Basal Cell Carcinoma	3323
Benign Keratosis	2624
Dermatofibroma	239
Melanocytic Nevus	12875
Melanoma	4522
Squamous Cell Carcinoma	628
Unknown	0
Vascular Lesion	253

training deeper neural nets, there is a chance of model overfitting, leading to decrease in validation accuracy. Due to this, we are also looking into data augmentation using GANs. GANs can be used to generate data, but have not been used by supervised training algorithms Pollastri et al. [2019]. In Federico Pollastri [2019], they improve GANs by making it possible to be used as additional training examples. They implemented an architecture that generated both the image and its segmentation mask, allowing for use of new images as additional training data. As the classes in our dataset are unbalanced, we looked into using GANs to generate more data for the imbalanced classes.

4 Dataset

Our dataset consists of the training and test data from the ISIC 2019 classification challenge, which was collected from multiple sources Tschandl et al. [2018], Codella et al. [2018], Combalia et al. [2019]. The training dataset consists of 25,311 different skin lesion images, and 8,238 test images. Each image from the training set is classified as one of several different types, which are listed in Table 1. These types vary from highly dangerous, fast spreading cancers such as Melanoma, slower spreading, much more common cancers such as Basal Cell Carcinoma, and noncancerous growths such as Benign Karatosis. Notably, in the dataset there also exists an unknown type. This is a type of lesion which exists in the test set, but no training data is provided for. Correctly classifying this class requires finding outliers in the test data that don't fall into any of the training classes.

All images vary in brightness, skin and lesion colors, lesion size and position. While color variation is expected in most image recognition algorithms, position variation is surprising, as it would be expected that images would be centered on the area of interest. While most images are properly centered, some images contain only small growth that are not centered. The difference in growth sizes in images could be due to the differences in the size of the growths or the amount of zoom applied to the image, which is another common problem for image classification. Overall, almost all images contain the full growth, with only a few cutoff around the edges.

The dataset also includes some variations that may not be common in all image classification data-sets. Some images are differently styled, as some are full square pictures of a region of skin, while other have been cropped into circular images with black borders surrounding the region of interest. Several images include patient hair over top of and around the skin lesions of interest, creating features that we may want to remove or train the classifier to ignore. Some images also include rulers, pads, or other devices used by physicians when analyzing the lesions as well. A few images include multiple growths, which could create some confusion about which area in the image is the most important to focus on, and some images have been added to the dataset multiple times, sometimes under different classifications. The size of the images also varies. The longest side is usually 1024 pixels with a smaller side of 1024 or 768 pixels, however both sides vary between 512 and 1024 pixels.

5 Experiments

Baseline

For a simple evaluation of different types of network, we ran six networks, ResNext, Densenet, and two versions of each Resnet and Efficientnet, and compared accuracy after five epochs using a small image size (224x224) and scoring on unweighted accuracy, instead of class balanced accuracy, with the results shown in 1 below. Overall, both EfficientNet models scored the highest, leading to us adopting it as our primary model for testing purposes. This initial experiment was run on networks without pre-trained weights, but we believe it is a good representation of how fast the models can adapt to the data we are using.

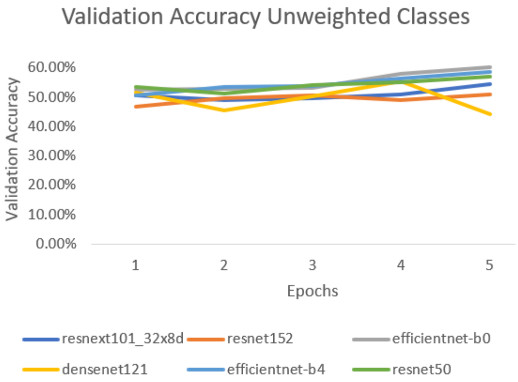
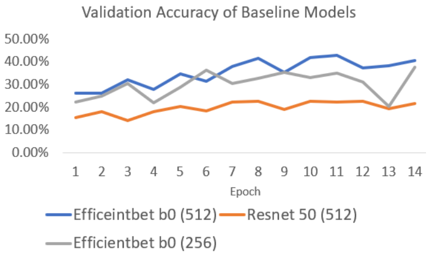


Figure 1: Validation accuracy over five epochs using size 224x224 images. The accuracy was not balanced for the number of images per class, which allows algorithms that successfully predict images from the largest classes to obtain a high accuracy (close to 60%). In this experiment, Efficientnet-b0 outperformed all other methods, and became our baseline model.

After the initial experiment, we calculated the baseline using balanced class accuracy as a metric on images with a final image size of 512x512. Efficientnet-b0 outscored the Resnet-50 model significantly, peaking at 42.3% accuracy. Using 256x256 images as the final size, the results for Efficientnet-b0 still did well, but the results were not as good as they were for the higher resolution images. Results for both can be seen in Figure 2.

Color Spaces

We tested the effect of running different color spaces on the network, as well as combining the four color spaces together. Our first attempt, using a convolution layer to collapse the 4 color spaces into 3 channels for use by the EfficientNet model performed the worst, as it added a layer to the top of the network, nullifying the benefit of having the pre-trained weights. Of the individual color schemes, RGB performed the best. This is more than likely due to using pre-trained models, which were most likely trained using RGB images. However, other color schemes, such as LUV and LAB,



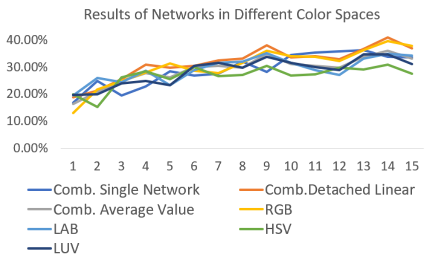


Figure 2 (Left): Results of three different baseline models run over 14 epochs. Overall, both efficient net models out performed Resnet by 15-25%

Figure 3 (Right): Results of seven different networks with different coloring schemes. Combining the four individual color space networks with a detached linear layer slightly outperforms RGB images (peaks of 40.86% and 39.55%), followed by the average of the four networks and the LAB network (peaks of 36.13% and 34.91%). Given that the network was pretrained most likely on RGB images, the fact that they yield even a slight improvement implies it may be useful in the future to boost results.

also perform well, but were unable to completely catch RGB images in 15 epochs. While RGB image do test better than the other color schemes, combing the results from of the four networks using a detached single layer network managed to slightly improve results beyond what RGB alone achieved. However, these improvements still fell slightly short of what was accomplished with 512x512 images, which required the same amount of computational effort. The results of these 7 experiments can be found in Figure 3.

U-Net

Focused lesion features can play a crucial role in classification of skin cancer type. We used the state-of the art segmentation technique for Biomedical Image Segmentation called U-net. In our pipeline we used segmentation of lesions and performed the further pre-processing on the segmented images. However, no improvements were observed on the overall accuracy.

Hair Removal

We tested the effect of removing hairs for networks with a final image size of 256x256. Overall, we found that the results of EfficientNet to be improved by over 1.5% when compared to the best run with no hair removal using 256x256 images as the final size. Images of the hair removal process can be see in figure 4, and the results of the network can be found in Figure 5.

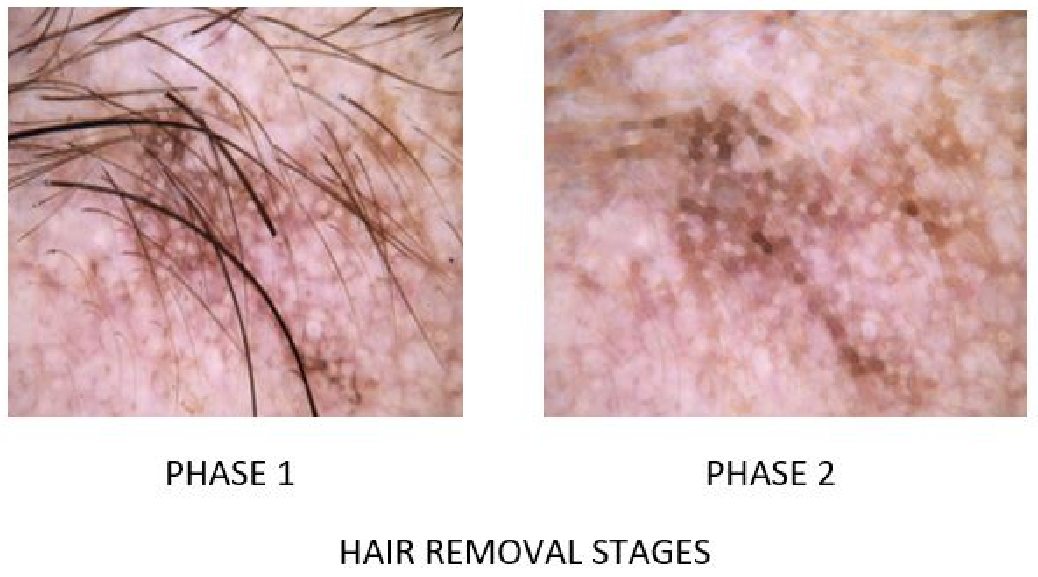


Figure 4: Phases of hair removal

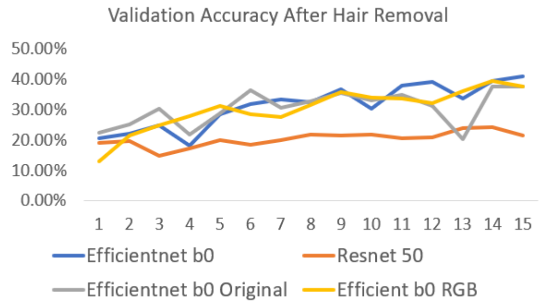


Figure 5: After removing hair, Efficientnet b0 outperforms both the original run and the RGB run by 1.5% - 4%, scoring 41.03% accuracy on its final epoch. Resnet-50 peaks around 25%, significantly lower than the other 3 models.

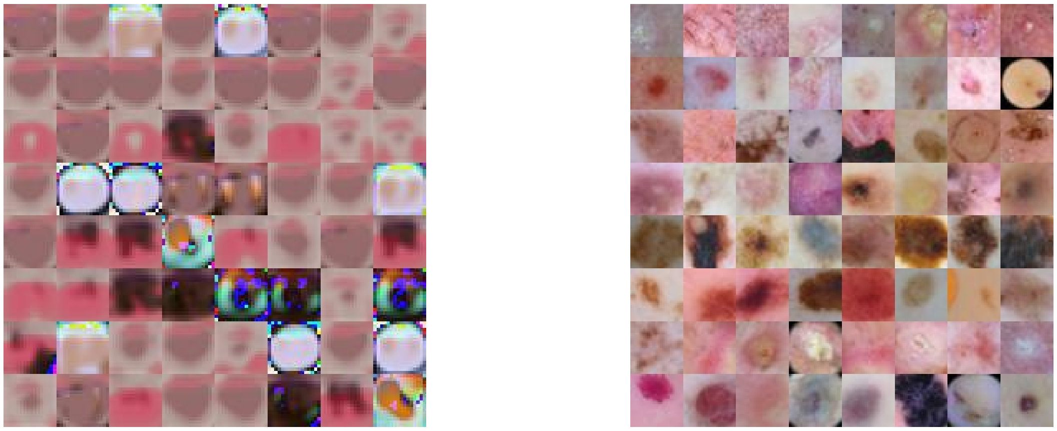


Figure 6 Images generated by GANs (left) versus images from the training data (right). Each row represents images from a different class. Overall, our GANs procedure wasn't converging everywhere (as seen in the multi-colored images) and didn't manage to map the minor variations of color on skin and lesion areas in a realistic manner. Several images appear similar in feature, but that could its uncertain it that is an early sign of mode collapse or just a sign of the network struggling to make realistic images. It is also noticeable that there are similar images on different rows (different classes), which is something we will look into further in the future.

Despite slight improvements from removing hair and ensemble color spaces over the baseline, class imbalance is still the main contributor to the low accuracy scores. Table 1 shows the individual class accuracy for some of the best models used on their best epoch. Overall, the five smallest classes were under 60% in all models and no classifier successfully predicted a single case of Dermatofibroma. No model even predicted any validation data as belonging to the Dermatofibroma class on its best epoch.

GANs

As shown in Table 1, class imbalance plays a large role in the overall accuracy of the model. To offset this, we tried to implement a GAN model to generate images. We used the start of a conditional GAN, with a latent size 100 and embedding size of 16, on top of the architecture for a progressive GAN Karras et al. [2017]. We started with a resolution size of 16, and attempted to double the size until reaching the resolution we needed. We chose this method based on the claim that progressive GANs can train a high-resolution image up to six times faster than a regular GAN. We also used Wasserstein GAN with Gradient Penalty (WGAN-GP) Gulrajani et al. [2017] as our primary loss function and used a linear activation instead of a sigmoid or tanh, based on the recommendations of the progressive GAN architecture. PGANs utilized an extra penalty to constraint the value of the critic to being close to zero, however, in our model, we instead tried to use pixel normalization in the discriminator to make it match the generator and to replace other normalization layers, such as batch normalization, which cannot work with WGAN-GP. This differs from the official pGAN implementation, which does not use normalization in most layers of the discriminator relying primarily on their extra penalty to normalize the network. Overall, we trained the network for three days, with the majority of the time spent on 32x32 resolution. However, loss never fully converged, and images never sharpened as much as we would have liked. A comparison between our GAN generated images and real images at the 32x32 resolution can be seen in figure 4.

6 Conclusion And Future Work

We explored Efficientnt-b0 and Resnet-50 with the ISIC skin cancer detection dataset, using different pre-processing methods to try to improve results. Our baseline model at 256x256 resolution images performed slightly below 40%, while two of the techniques we utilized, ensemble of different colored network and hair removal, both tested above the 40% mark, showing a slightly improvement over the baseline. It should be noted that our baseline model at 512x512 resolution had a balanced class accuracy of 42%, however, we believe that our method would improve the performance of the model at that resolution as well, and could perform even better if both were used at the same time.

GANs were also explored for data augmentation. Although we were able to produce some generic images, fine levels of detail that would make them realistic couldn't be achieved. Our GANs model produced near solid color lesions and skins, while real images showed variations in the coloring. We believe this was due to various reasons, ranging from the latent size to the difficulties in the structure of the images. The progressive GANs in Karras et al. [2017] used a 512 latent vector, whereas we used a 100 dimension latent vector with an additional 16 dimensions from an embedding (class) vector. Also, there was no general construct for spatial positioning of elements as the images can have different colored elements in different positions. This could possibly be a reason for why our algorithm struggled even more from pulling away from a generic skin color, in addition to the extra work it already had in generating eight classes instead of one. In the future, we will look into GANs to augment data for the minority classes.

We also experimented with U-Nets for segmenting out lesion from the images. They failed to improve the model accuracy because of the possibilities of multiple lesions in a single image and high similarity levels of lesions and surrounding pixels. For future work, localization techniques can be employed before segmentation, focusing on major lesion areas. Other state-of-the art segmentation algorithms like GrabCut can also be explored. Using contrast enhancement techniques can also make significant difference in the segmentation of lesions. Also hardware constraints hindered the training of our models for higher resolution images.

7 Acknowledgements

We would like to thank Professor Bhiksha Raj, our project mentor Hanna Farooq Moazam, and the rest of the 11-785 staff for all their support and mentorship.

References

Devansh Bisla, Anna Choromanska, Russell S Berman, Jennifer A Stein, and David Polsky. Towards automated melanoma detection with deep learning: Data purification and augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan C Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

Mario Parreno Federico Bolelli Roberto Paredes Costantino Grana1 Alberto Albiol Federico Pollastri, Juan Maronas. AImageLab-PRHLT at ISIC Challenge 2019. Technical report, Universit'a degli Studi di Modena e Reggio Emilia and Universitat Polit'ecnica de Val'encia, Italy and Valencia, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL `http://arxiv.org/abs/1710.10196`.

Federico Pollastri, Federico Bolelli, Roberto Paredes, and Costantino Grana. Augmenting data with gans to segment melanoma skin lesions. *Multimedia Tools and Applications*, pages 1–18, 2019.

Philippe Schmid-Saugeona, Joël Guillodb, and Jean-Philippe Thirana. Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics*, 27(1): 65–78, 2003.

Rusong Meng Steven Zhou, Yixin Zhuang. Multi-Category Skin Lesion Diagnosis Using Dermoscopy Images and Deep CNN Ensembles. Technical report, 2019.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.