Project H40

Non-RRI

Using a Generative Adversarial Network and an ensemble of Convolutional Neural Networks to More Accurately Generate and Diagnose Skin Condition Datasets in Diverse Skin Types

All 11th Grade

Johnathan Mo, Grant Sims, Robert Zhang

Introduction

In the approximate 19 million cancers that are diagnosed each year, one in every three cancers diagnosed is a skin cancer, with ~1.9 billion people worldwide suffering from some form of skin condition. Many of these conditions have visually similar symptoms but present differently on different types of skin, specifically darker skin tones. This inconsistency leads to misdiagnosis because most practicing dermatologists are trained on lighter skin.

Furthermore, the small number of practicing dermatologists means that general practitioners often see up to half of skin-related cases. In fact, the predictive diagnostic accuracy of general practitioners is as low as 24% while that of dermatologists is as low as 77%, leading to inaccurate diagnoses, delays in care, and errors in treatment. This issue is compounded by the fact that existing datasets lack representation among darker skin tones, resulting in a decrease in dermatological diagnosis accuracy among existing machine learning models.

Fortunately, the rise of machine learning has allowed for the creation of advanced neural network models that are capable of assisting medical professionals in the diagnosis of various medical conditions. Furthermore, the development of Generative Adversarial Networks has allowed for the generation of completely new images that are similar in appearance to real-life images.

In our project, we used generative adversarial networks and Convolutional Neural Networks to generate and classify dark-skin images.

Rising Case Numbers

Currently ~3 million non-melanoma skin cancers and 132,000 melanoma skin cancers occur globally each year, a 44% rise in cases over the past decade.

Diagnostic Bias

These late, inaccurate diagnostics lead to inadequate treatments, and according to a study from the AAD, African Americans have the highest mortality rate for skin cancer, having a lower 73% five-year survival rate compared to 90% for light-skinned Americans.

Lack of Data

In an MIT study of several dermatology atlases, there were 3.6 times more images of the two lightest skin types than the two darkest skin types. Additionally, the study found that an average of only 89 represented diseases in darker skin types versus the total 114 skin conditions in lighter skin types.

Conditions

Out of the two major types of skin cancer, Melanoma is fatal and has an estimated five-year survival rate of about 99% if detected early and 20% if detected late. In the year 2018, there was an estimated 96,480 new cases of melanoma and 7,230 deaths. Additionally, Purpura is on the rise with significant jumps in diagnosis numbers over the past few years.

Objectives

To expedite the diagnosis process and improve accuracy, healthcare professionals are turning to automated systems to aid dermatologists in their diagnostics with the use of machine learning models, which have been proven to increase average diagnostic accuracy by up to 63% in practitioners and 7% in dermatologists. These models are trained using compiled datasets of patient images. However, existing datasets lack representation among darker skin tones, resulting in a decrease in dermatological diagnosis accuracy of these models on darker-skinned images. The larger amount of lighter-colored skin datasets available for dermatologists and the more common use of lighter-colored skin for dermatological studies and training make these machine learning systems inadequate for use on darker-skinned individuals. Our project aims to address the lack of darker skin image data for Melanoma and Purpura and the lower diagnostic accuracy on darker skin images by producing more medically accurate darkskin images that can be used on future models for accurate and consistent results.

Quantity

The impact of our project is magnified by the number of successful images we can generate and ultimately spread to skin condition datasets and increase dark-skin representation.

Quality

While quantity is important, the quality of our generated images is just as important, as we want to make sure the images are an accurate representation of an actual medical image.

Accuracy

To validate our generated images' usability in other datasets and training, we must ensure that our images contribute to the training of a model. Thus, we will train a network of classifiers on generated images and test them on original skin images.

Scalability

Our generator must be able to consistently generate accurate images for any skin condition provided suitable training data. Our network of classifiers should be able to train on and diagnose any skin image data given an adequate dataset

Algorithms

GAN

First implemented in 2014 by Ian Goodfellow, a generative adversarial network or GAN is an unsupervised machine learning framework consisting of two neural networks, a discriminator and a generator, which operate in tandem to detect patterns and features in input data in order to generate new examples similar to the original dataset.

CNN

A CNN is a type of neural network classifier that uses convolution to extract specific features from each image, much like the human brain does. CNN offers performance benefits over classical neural networks because a CNN is able to compress images while preserving the important image details to improve performance.

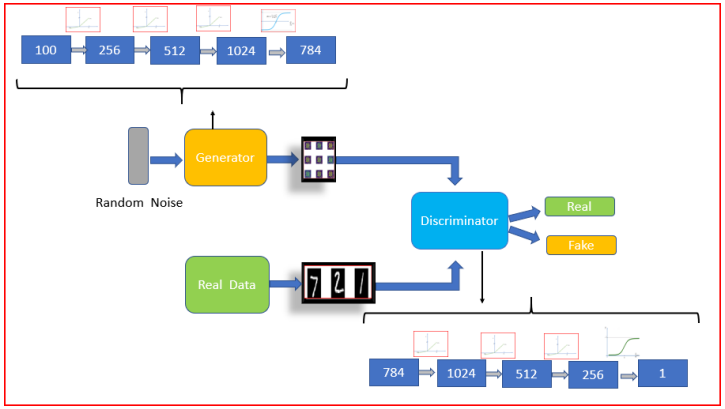


Figure 1: Architecture flowchart for a generative adversarial network (GAN).

Previous Work in Field

DermGAN

Synthetic Generation of Clinical Skin Images with Pathology

- Briefly mentions a range of skin tones in their dataset but does not include figures that indicate representation of darker skin tones in both inputted and generated images
- Uses adaptation of Pix2Pix GAN architecture
- Accuracy was decreased in some cases with the addition of the GAN
- Added 20,000 synthetic images to original training data of 49920 images.
- The overall performance is comparable to the baseline, but the performance on rare conditions like Melanoma and Basal cell carcinoma has noticeable improvement.

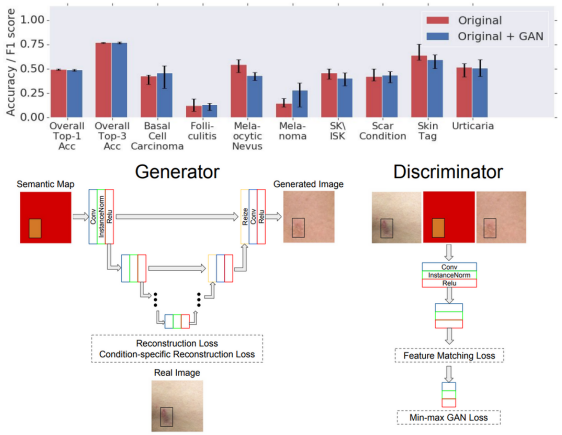


Figure 2 (top): Accuracy & F1 Score for DermGAN network with and without GAN.
Figure 3 (bottom): DermGAN network architecture flowchart.

University of Toledo

Skin Cancer Detection using Generative Adversarial Network and an Ensemble of deep Convolutional Neural Networks

- Classifies melanoma and benign skin lesions with ~84% accuracy (#3), but overfit on models #1 and #2.
- 3 Models
  - #1: Used DeBlur GAN and traditional data augmentation on grayscale images
  - #2: Used cGANs with ResNet-50 classifier on RGB images
  - #3: Used DCGAN and novel CNN ensemble classifier on RGB images
- Similar sensitivity & specificity, low accuracy among models due to differences in input image color space
- Used selection of conditions from HAM10000, which is a majority light-skin dataset and one of the most widely used for skin condition training.

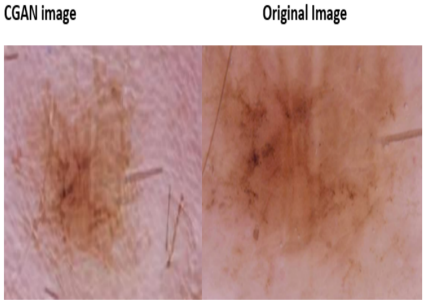


Figure 4: Original and generated image using cGAN network of Model 2 in this study.

CAMP - Johns Hopkins

Generating Highly Realistic Images of Skin Lesions with GANs

- Used ISIC-2018 Dataset
  - 99% light skin data
- Uses ImageNet-pre-trained models
- LAPGAN (Laplacian GAN)
  - Progresses up image size using upscaled generated images as a conditional comparison for the next resolution step
- DCGAN (Deep Convolutional GAN)
  - Progresses upwards through convolution and transposition layers until output resolution
- pGAN/ProGAN (Progressive GAN)
  - Downscales images then trains on increasing resolutions for a high-quality image result
- Evaluated the quality of the PGAN samples with expert dermatologists and Deep Learning experts
- Trained ignoring presence of different types of skin tone

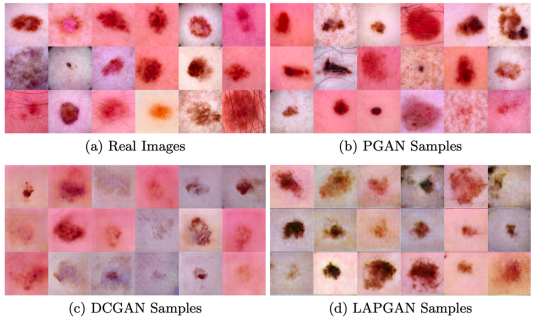


Figure 5: Chart of real and generated images with different GAN models used in this study

Data Acquisition & Flow

Source Images

VisualDX is a web-based clinical decision support system that is used in clinical applications to help physicians diagnose and treat conditions in dermatology. VisualDX claims that 28.5 percent of images are of individuals with skin tones that fall within levels IV-VI of the Fitzpatrick scale, which means they are of darker complexion [2].

Figure 6: Darkskin ground truth (real) images of Melanoma(left two images) and Purpura(right two images) sourced from VisualDx

Image Organization

We download dark-skin condition images such as the ones shown in Figure 6 from VisualDx for Melanoma and Purpura and upload them to Google Drive into a parent 'data' folder with separate training and testing labels.

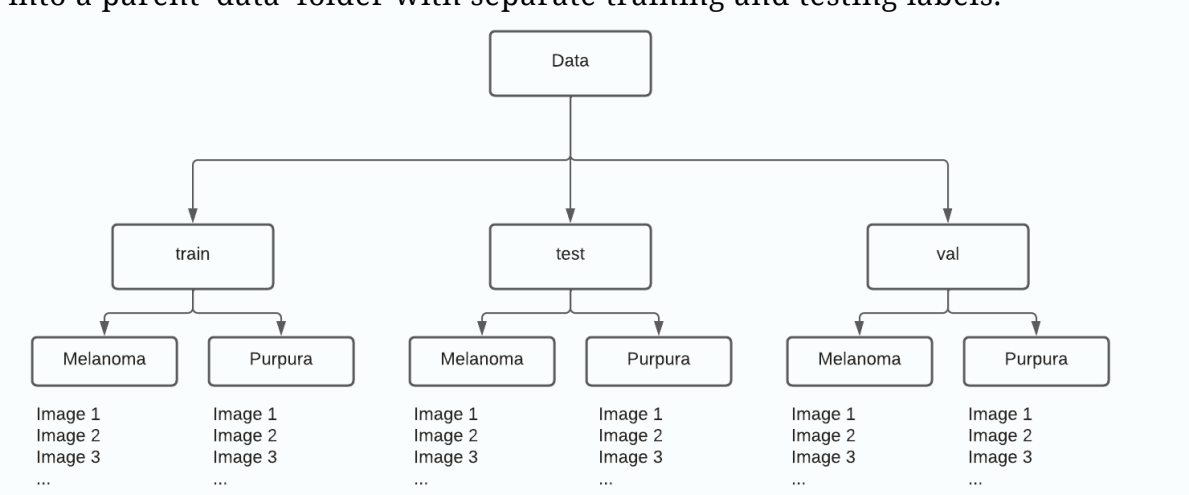


Figure 7: Image of our data organization folder hierarchy and data separation portions for training, testing, and validation.

Source Images

We source our images from VisualDx, an online dermatology atlas with a number of dark-skin images.

Template Input Generation

Using a novel feature detection algorithm, we generate noise and feature template inputs for each real image. The images are then loaded into template image folders for the condition.

Image Generation & Training

The random noise or feature images are passed into our GAN to generate an image based on the noise or feature template image. We compare the generated image with the original ground truth image for GAN training.

Image Organization

We upload the generated images to folders in Google Drive, from which we can easily access files from Google Colaboratory. We split up the images into 80% train and 20% test for each condition to prevent overfitting.

Classfication & Training

Generated images are resized and reformatted between each model to train the CNN ensemble.

Testing

Test ground truth images are used to test the CNN ensemble. If metric thresholds are not been met, we will retrain.

Data Acquisition & Flow

Template Image Generation

Each ground truth (real) image that we pulled from VisualDx then has a corresponding template noise image (phase 1) or a feature template image (phase 2) generated for it in order to train the GAN. After the GAN has been trained, we use the novel noise generator (phase 1) and template feature detector (phase 2) to generate a number of random noise or template images to train the GAN for each condition.

Generation and Training

The template images are fed along with their real image counterparts to the training loop of the GAN, which trains the GAN to generate realistic skin conditions given an input noise or feature template image. When a certain image quality threshold has been reached, 500 images are generated for each condition and separated into training and testing data folders as shown in Figure 7.

Classification and Training

The CNN Ensemble then trains on these training and testing images, and the quality of each model in the ensemble is measured by running the model on real validation images and calculating relevant metrics.

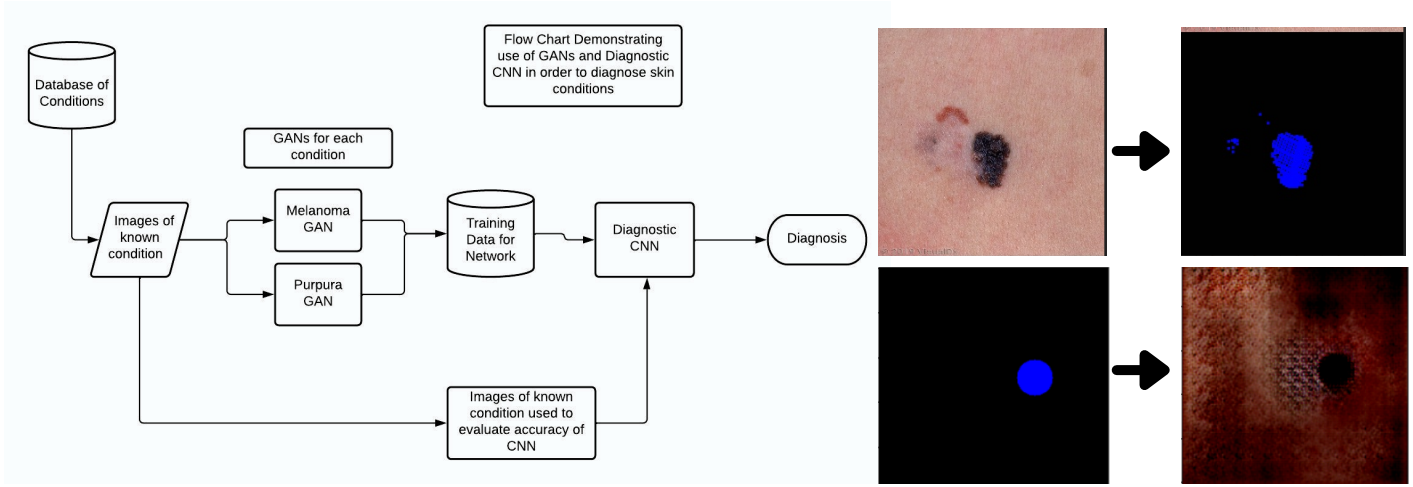


Figure 8: Flowchart of image data flow through GAN and diagnostic CNN.

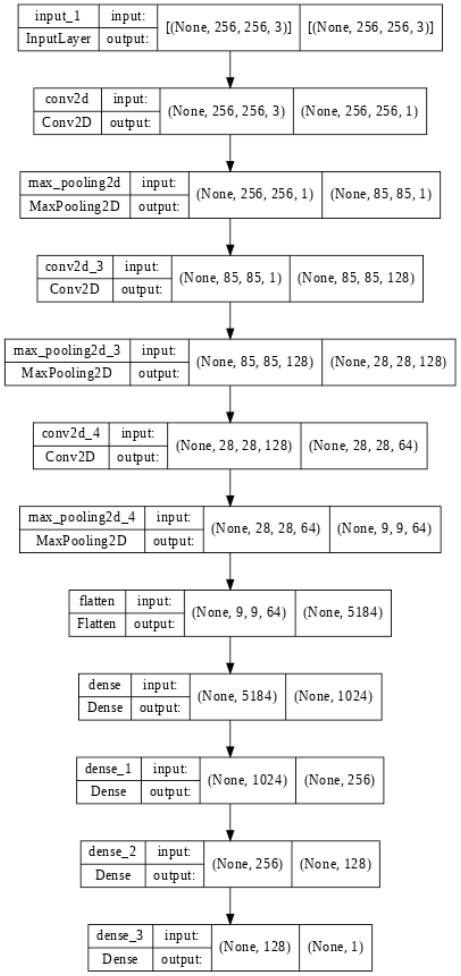


Figure 9(above): Diagram of our custom CNN Architecture with layer input parameters and outputs

Figure 10: Original ground truth and template feature image generated from ground truth image.

Figure 11: Template feature image and generated skin condition image using phase 1 GAN.

Methods - Phase 1

In phase 1, we used one GAN and one CNN to test our general project framework on Melanoma only.

Our steps included:

- Creating noise templates for each image
- Evaluating noise templates to generate similar images based on noise template
- Classifying images into given conditions (Melanoma vs Non-Melanoma) to make a diagnosis

Elements:

- Noise Generator
- GAN: Generative Adversarial Network
  - using novel Pix2Pix architecture
- CNN: Convolutional Neural Network
  - computationally efficient, quick feature detection, decent accuracy

CNN Metric: Evaluating a Confusion Matrix

A *true positive* is a case where the model correctly predicts the class for the condition being evaluated or positive class. On the other hand, a *true negative* is when the model correctly predicts the negative class. A *false positive* is when the model incorrectly predicts the positive class when the true class is negative, and a *false negative* is when the model incorrectly predicts the negative class when the true class is positive.

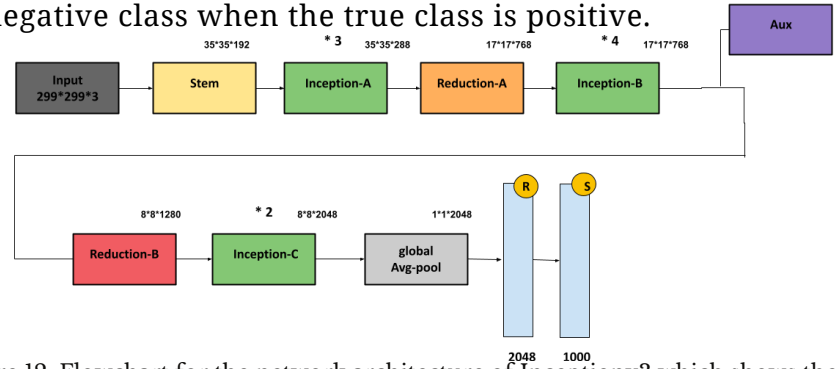


Figure 12: Flowchart for the network architecture of Inceptionv3 which shows the different processing methods and functions used in order from input to output.

Noise Test

We used a random noise generator function to create noise for each real image, hoping that the GAN would be able to generate features based on random noise.

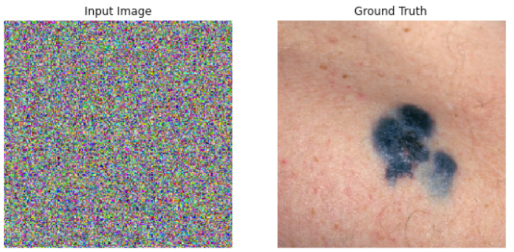


Figure 13: Ground truth image and Phase 1 noise generator-generated noise image

Generative Adversarial Network (GAN)

For our generative adversarial network (GAN) model, we used a novel algorithm based on the Pix2Pix architecture with two-dimensional convolutions and LeakyRelu activation functions in our generator and discriminator.

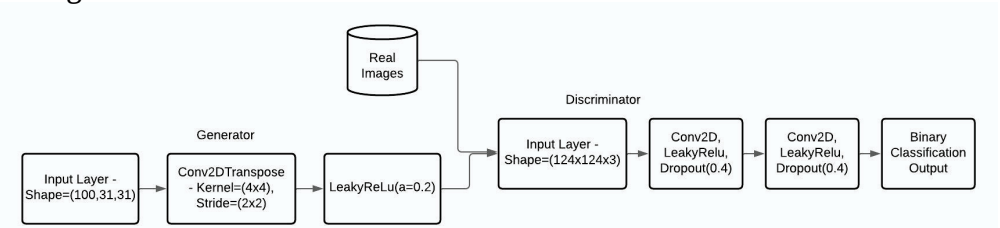


Figure 14: Flowchart for the network architecture of Inception which shows the process of image inputting, what steps are taken between layers, and the output.

Convolutional Neural Network (CNN)

We used Inceptionv3 as our CNN model for training on large datasets including skin images. To save time and resources from having to train multiple machine learning models from scratch to complete similar tasks, we used transfer learning, which retrains a high fidelity pre-trained model on our dataset.

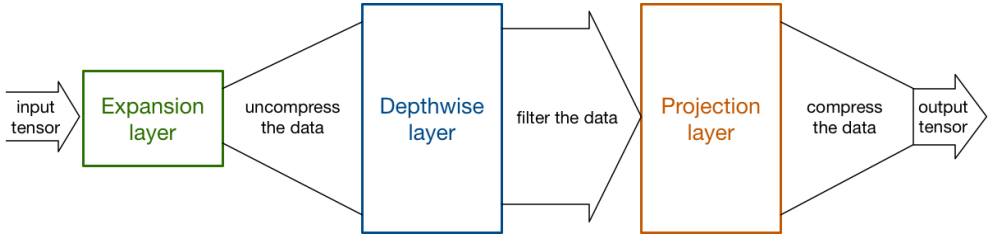


Figure 15: This is a flowchart for the network architecture of Inception which shows the process of image inputting, what steps are taken between layers, and the output.

Phase 1 Results

CNN | Confusion Matrix Heat Map

Melanoma

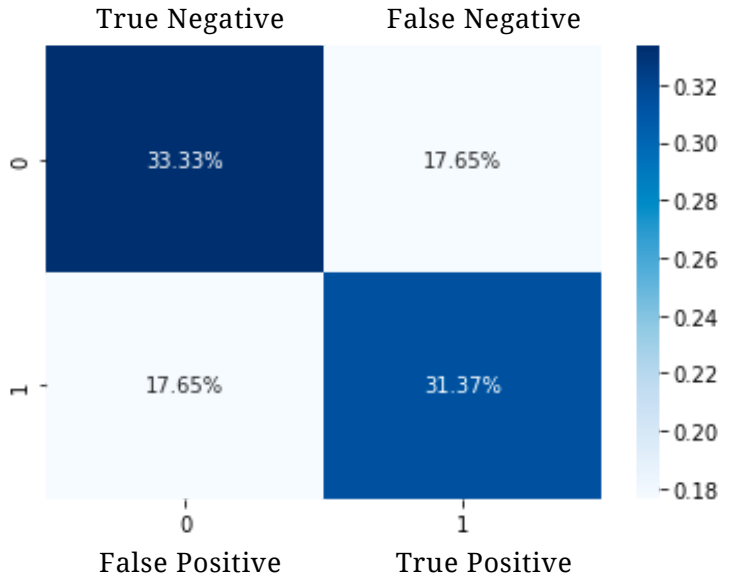
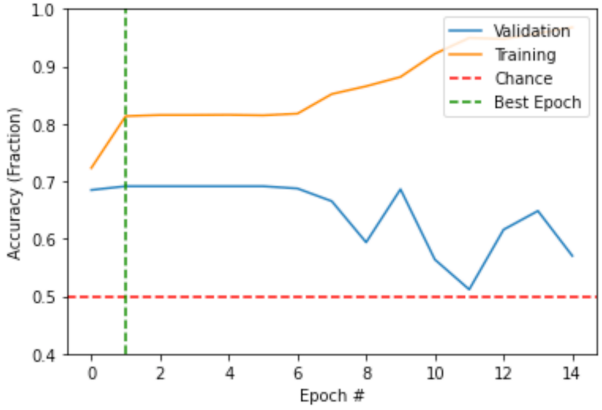


Figure 16: Confusion matrix representing True Negatives, False Negatives, False Positives, and True Positives for our custom Melanoma CNN.



f1 score: 0.64
recall: 0.64
accuracy: 0.647059
precision: 0.64

Figure 17: Plot of Validation Accuracy (Blue), Training Accuracy (Yellow) across each epoch, a portion of the dataset. The red line represents a guessing accuracy of 50%. The green line marks the epoch with the highest accuracies.

GAN Results

Figure 18: Noise input image #1 generated from ground truth and the predicted image from noise template.

Figure 19: Noise input image #2 generated from ground truth and the predicted image from noise template.

Discussion

From the GAN results, the images are clearly artificially generated but possess a strong resemblance to real melanoma images that the GAN was trained on. As shown in Figures 18 and 19, the same ground truth with different noise images can yield different predicted images. Looking at the CNN Confusion Matrix results in Figure 16, we see an f1, recall, accuracy, and precision score all in the vicinity of 64%, showing that our CNN trained well purely on the images generated by our GAN. Based on the results, we determined that the noise images were bottlenecking the success of the GAN and switched to a novel feature detection method to generate better template images, which led to better results in phase 2.

Methods - Phase 2

In phase 2, we drew insights from our results from phase 1 and implemented a few key changes:

- We replaced our random noise generator with a novel feature detector that generates image-specific template images to better train the GAN.
- We used a CNN Ensemble consisting of 4 CNN models: Inceptionv3, ResNet50, DenseNet121, and our own custom model

Elements:

- Feature Detection Algorithm
- GAN: Generative Adversarial Network
  - using novel Pix2Pix architecture
- CNN Ensemble
  - Testing 4 different models for composite metrics

GAN Metric: Evaluating GAN Loss

While GANs do not have a solitary loss function as CNNs do, they can still optimize by means of a loss function, which evaluates how much the output images deviate from actual results. Larger discrepancies between the two mean a larger loss number while small discrepancies result in a low loss value.

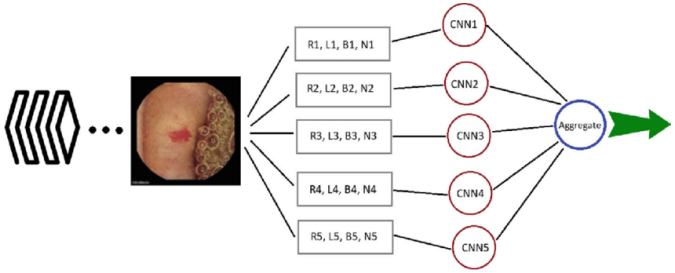


Figure 20: Flowchart for our implement-ation of a CNN Ensemble connected via ensemble learning.

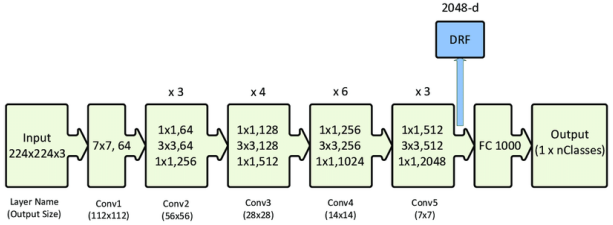


Figure 21: This is a flowchart for the network architecture of ResNet50.

Novel Feature Detection

We implemented a novel feature detection algorithm using algorithmic object detection to generate feature template images for our GAN to train on.

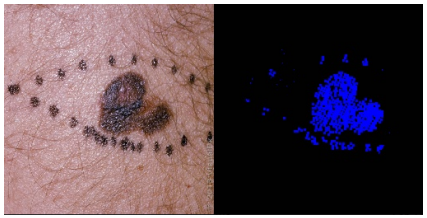


Figure 22: Ground truth image, Phase 2 feature detector-generated template image

Generative Adversarial Network (GAN)

For our phase 2 GAN model, we used the pixel-to-pixel or Pix2Pix GAN architecture. Conditional GANs including Pix2Pix use a convolutional network model called U-Net to segment images, convoluting the image with an activation function down into smaller sections until a certain size is reached, from which point the image will be convoluted upwards, raising the size of each section until the resolution of the original image has been reached.

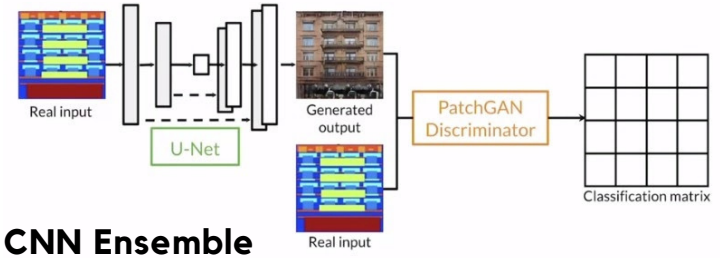


Figure 23: Pix2Pix Architecture for GAN consisting of the U-net based generator and PatchGAN based discriminator.

CNN Ensemble

We used a variety of Keras-based models in our ensemble, which included Inceptionv3, ResNet50, DenseNet121, and our custom model. In order to connect all of the models together, we used ensemble learning. Due to the smaller number of dark-skin images currently available, we used ensemble learning to exploit available training data by forming a final diagnosis output by merging the output results of each CNN, which eliminated network-specific biases and increased output classification accuracy compared to individual model accuracies.

Phase 2 Results

GAN

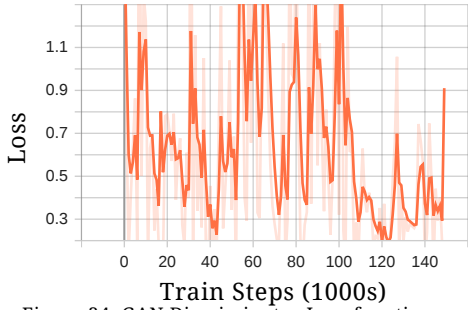


Figure 24: GAN Discriminator Loss function graph.

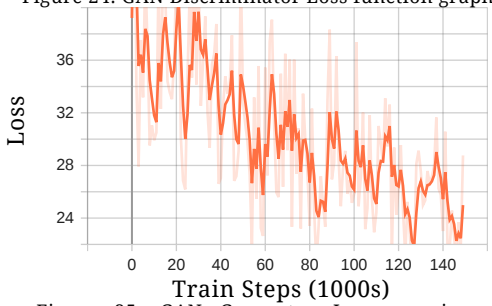


Figure 25: GAN Generator Loss per image function graph.

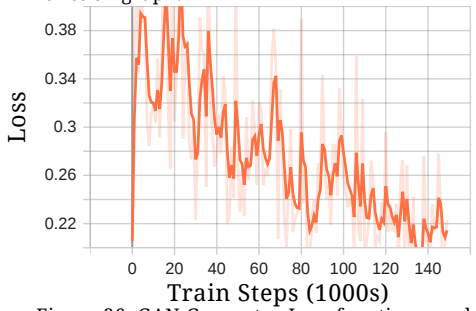


Figure 26: GAN Generator Loss function graph (all images)

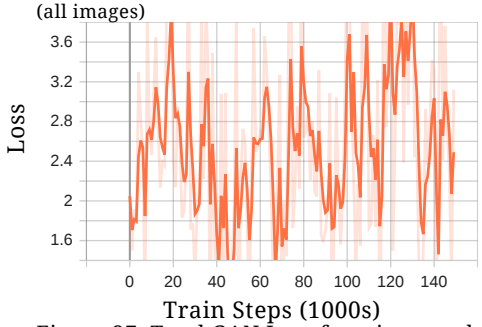


Figure 27: Total GAN Loss function graph.

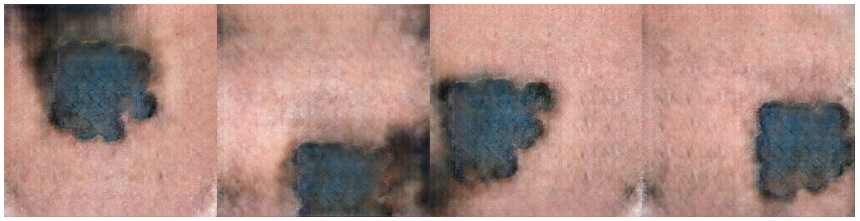


Figure 28: An array of generated melanoma images from our GAN using phase 2 random generated feature detection template images.

Figure 29: Ground truth image, Phase 2 feature detector-generated template image, and GAN generated melanoma training image on template image.

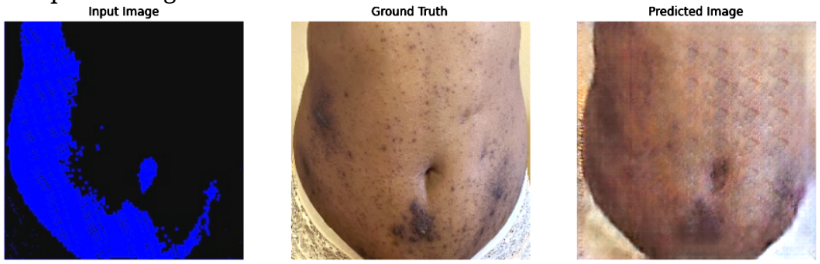


Figure 30: Ground truth image, Phase 2 feature detector-generated template image, and GAN generated Purpura training image on template image.

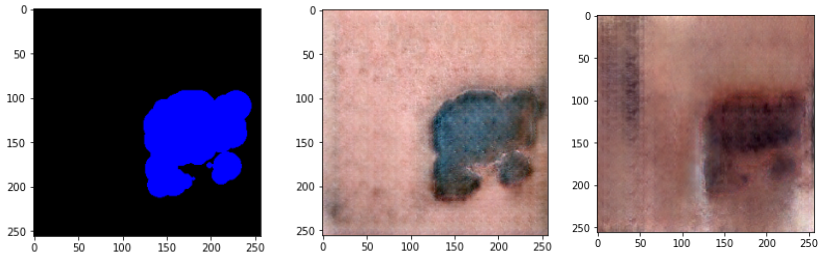


Figure 31: Randomly generated feature detector template image and generated images using Melanoma GAN (middle) and Purpura GAN(right).

Discussion

The discriminator and generator losses are loss metrics from the two parts of the GAN that generate images and evaluate their accuracy. The graphs, denoted by Figures 24 through 27, show that the GAN trained effectively. Indicated by our discriminator loss graph which varied from 0.3 to 1.1 and had an average of ~0.693, this average loss value means that the discriminator could not differentiate the GAN's generated images from real images. Furthermore, as shown in Figures 28 - 31, the generated images clearly look more similar to real skin images when compared with the GAN from phase 1. The efficacy of the template images compared to the input images in combination with the improved GAN showed a clearer and more accurate diagnostic shape of the affected area and more accurate colors compared to the ground truth image. These improvements were also reflected in the CNN training results.

Phase 2 Results

CNN Ensemble

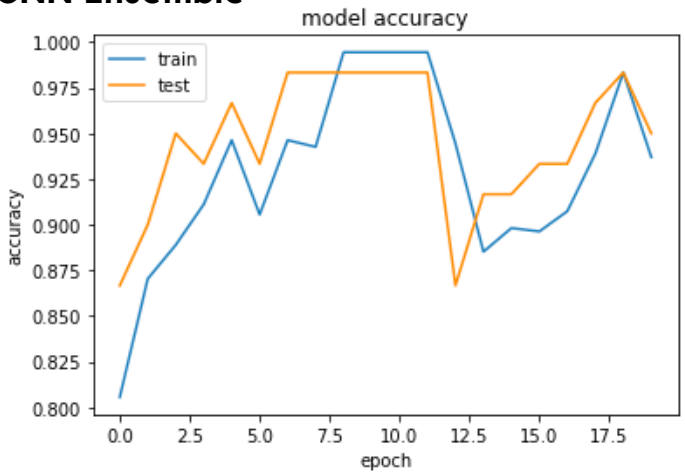


Figure 32: Plot of Train Accuracy (Blue), Validation Accuracy (Yellow) across each epoch. The Y axis is the accuracy of the model and the X axis is the epoch #.

Condition	F1 Score	Recall	Precision	Accuracy
Melanoma	0.86	0.84	0.88	0.86
Purpura	0.73	0.72	0.75	0.75

Discussion

As shown in Figure 32, the validation and training accuracies were highly consistent, showing that the CNN ensemble did not overfit like in phase 1.

Confusion Matrices

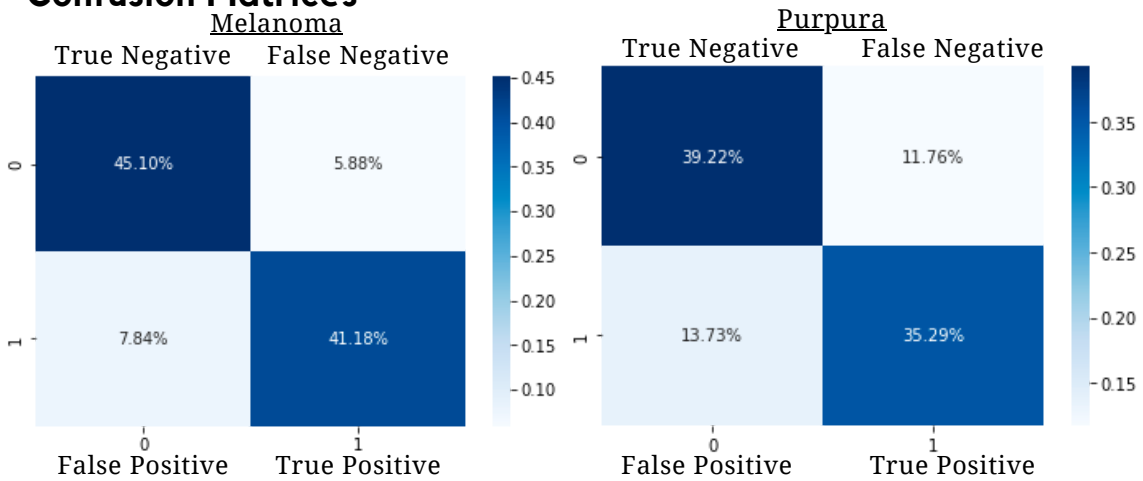


Figure 33, 34: Confusion matrices for Melanoma (top, 28) and Purpura (bottom, 29) representing True Negatives (top left), False Negatives (top right), False Positives (bottom left), and True Positives (bottom right) for our custom Melanoma CNN.

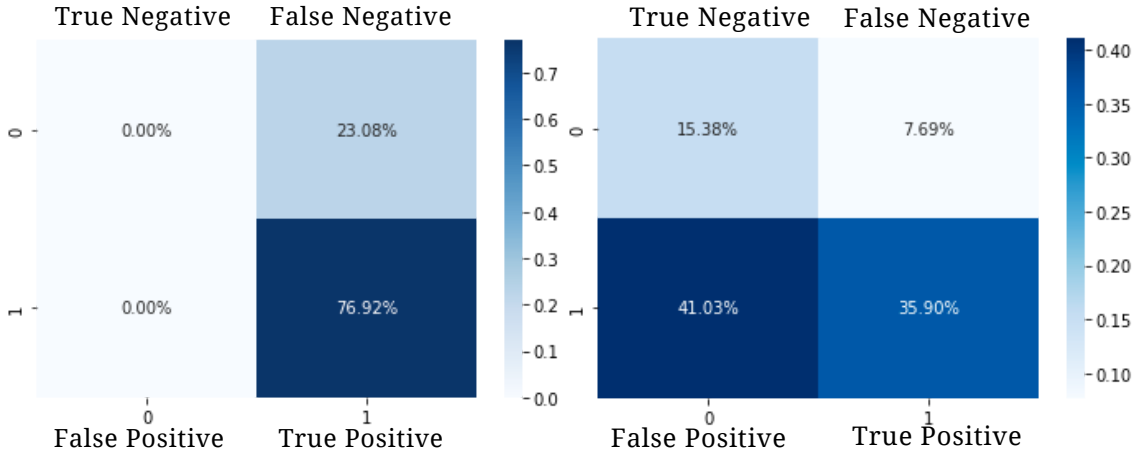


Figure 35, 36: Confusion matrices for Melanoma (top, 28) and Purpura (bottom, 29) representing True Negatives (top left), False Negatives (top right), False Positives (bottom left), and True Positives (bottom right) for a CNN trained solely on available real images (excluding GAN-generated images).

The confusion matrices shown in Figures 33 & 34 show higher true negative and true positive rates, with lower false positive and false negative rates. With Melanoma scores around 85% in f1, recall, accuracy, and precision, and Purpura scores around 74% in f1, recall, accuracy, and precision, the models prove to be remarkably accurate on field data. Furthermore, as shown in Figures 35 and 36, our GAN-generated images significantly improved CNN performance vs training solely on available field data images.

Conclusion

For our phase 1 results, while the GAN generated images are clearly artificial, they possess a strong resemblance to real melanoma images that they were trained on. Our CNN Confusion matrix heat map for evaluation showed an f1, recall, accuracy, and precision score all in the vicinity of 64%, meaning our CNN trained decently purely on the images generated by our GAN. In phase 2, our novel feature detection algorithm, modified GAN, and new CNN ensemble proved to increase our metrics significantly, with confusion matrix metrics all-around 85% for Melanoma and 74% for Purpura. Thus, the models proved to be fairly accurate on field data.

Applications

Overall, our results demonstrate that our GAN is able to produce images that accurately mimic the diagnostic characteristics of Melanoma and Purpura, two conditions that are present differently in dark skin individuals and are often diagnosed at a less frequent rate. Furthermore, we have successfully used this data to train a CNN Ensemble to differentiate two similar skin conditions. Applications of our work include using GANs to generate accurate medical training images for dermatologists using limited available data. The implications of this work include reducing inequality for darker skin individuals who lack representation in medical training images as well as providing a diagnostic tool to aid dermatologists in diagnosing skin conditions.

Sources of Error

Overfitting:

Overfitting is when a model essentially begins to memorize the specific training set, causing training accuracies to soar while validation accuracies fluctuate wildly. Overfitting is present in the CNN of phase 1. Some ways to combat overfitting include introducing dropout layers into the model which force the model to generalize, increasing the amount of data used, or altering the learning rate to decrease the network's tendency to overfit

Generalization Error:

As shown by our fluctuations in phase 2 Generator Loss, there is a high variance in generalization, meaning that the generator is not efficiently using its past generalizations on the newer images. Solutions for the generalization error include altering the structure of the GAN model or utilizing a different base model architecture, such as CycleGAN.

Limitations

Extraneous Features:

Certain images have shadows or dark areas that aren't due to features we want to target, but are detected as such, which decreases the accuracy of the detector and may harm GAN training performance.

A solution for the extraneous features problem is creating a radius around which features would be ignored, as well as calculating the eccentricity and smoothness of the features to differentiate actual from noise.

Future Research

We would like to see the metrics of different GAN architectures in the future including ProGAN, DCGAN, and CycleGAN compared to our own GAN. Additionally, training on and evaluating results for more conditions with higher resolution images would be insightful into the scalability of our GAN.

Works Cited:

[1] https://arxiv.org/abs/1611.07004
[2] https://arxiv.org/abs/1911.08716

Summary:

Despite our identified sources of error, we were still able to create a novel feature detection algorithm that generated highly compatible template images for our GAN, which in turn was able to generate accurate and quality images of Melanoma and Purpura. Finally, our CNN ensemble including our own custom model was able to train on our GAN generated images and classify new ground truth images with higher accuracy than both general practitioners and dermatologists.