A Novel Approach to Promote Equity in Skin Disease Diagnosis by AI Models

Varsha Narasiman

American High School

November 6, 2023

A Novel Approach to Promote Equity in Skin Disease Diagnosis by AI Models

Abstract

AI-based systems are increasingly used to diagnose skin diseases with datasets available on the internet. However, the training data predominantly represents fair skinned people. The goal is to increase the accuracy of existing models in diagnosing skin disease across various skin tones within 10% of that obtained in diagnosing fairer skin tones, which is about 95.8%. The publicly available HAM10000 dataset with 10000 fair skin images was preprocessed and used to fine-tune a Vision Transformer model trained on the ImageNet-21K dataset. The model obtained about 94.9% accuracy in diagnosing fair skin images and about 19.8% accuracy in diagnosing real non-fair skin images. Next, 50% of the training data was transformed into non-fair skin simulation by CycleGAN based on 320 real non-fair images. After several iterations, the model achieved about 78.6% accuracy in classifying any skin tone and 52.1% accuracy in classifying real non-fair images. Though this 68.4% is outside of the desired range, it still falls within the accuracy of dermatologists (48–77%). However, it is the accuracy in classifying non-fair skin that makes this model significant. The model achieved a 28.6% increase in accuracy in diagnosing non-fair images using simulated data thereby narrowing the disparity. A confusion matrix was plotted to visualize the validity of the predictions. The functionality of this model suggests that similar data augmentation techniques could be applied to other AI models to ensure their fairness to all categories of people and correct any biases in data due to historic under-representation.

A Novel Approach to Promote Equity in Skin Disease Diagnosis by AI Models

The goal of this research is to promote equity in skin disease diagnosed by Artificial Intelligence algorithms. Datasets containing images of diseased and healthy skin of all skin tones have been utilized. Due to scarce availability of dark skin tones, generative models like CycleGAN will be used as supplements.

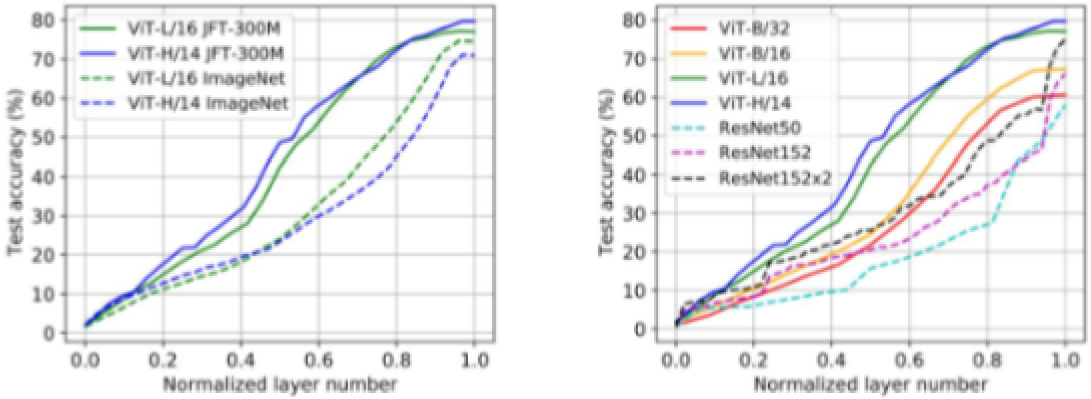
Introduction

Skin disease is one of the most common ailments affecting Americans, with 84.5 million or one in four being affected by some skin conditions according to the American Academy of Dermatology Association. Skin care costs about $75 billion per year, but rural populations still do not have access to dermatologists ("Burden Of Skin Disease"). In addition to access, care is also adversely affected by skin tone. Though fair skinned people are 3.33 times more likely to develop melanoma when compared to non-fair skinned people (encompassing Hispanics, Blacks, and other races), the latter group's fatality rate is 23% higher due to failure to detect the disease early, as stated in the report by Centers for Disease Control and Prevention (Balch).

Skin disease diagnosis is tricky even for experienced dermatologists, whose accuracy ranges from 48 to 77%, and AI models are extensively used to aid them (Yotsu et al.). However, due to the lack of real non-fair skin images, these models are trained only on fair skin types, resulting in wide accuracy gaps between skin tones. One of the few datasets to have exclusively non-fair images, the Dermatologically Diverse Dataset (DDI) by Stanford contains only 656 images compared to the 10,000 fair skin images publicly available in the HAM10000 dataset. As this is too small to be used in the training data of AI models directly, a carefully selected subset of the Fitzpatrick 17k dataset was used. Accuracy of some labels in this dataset is not fully verified by dermatologists, so only the verified images were used.

Related Work

The research paper titled "Multi-Class Skin Problem Classification Using Deep Generative Adversarial Network (DGAN)" by Maleika Heenaye-Mamode Khan, Nuzhah Gooda Sahib-Kaudeer, Motean Dayalen, Faadil Mahomedaly, Ganesh R. Sinha, Kapil Kumar Nagwanshi, and Amelia Taylor, published in the *National Library of Medicine* in March 2022, used a GAN model, a generative AI family in which CycleGAN belongs. It was also one of the first papers to include both fair and non-fair skin images, and consequently, had to deal with the problem of non standard data. The technique of enhancing an image to increase the contrast, which makes it easier for the model to find the lesions, was innovative. While the paper was oriented towards CNN models, this research leverages the potential of the Vision Transformer, as it requires less computational power to reach the desired level of accuracy.



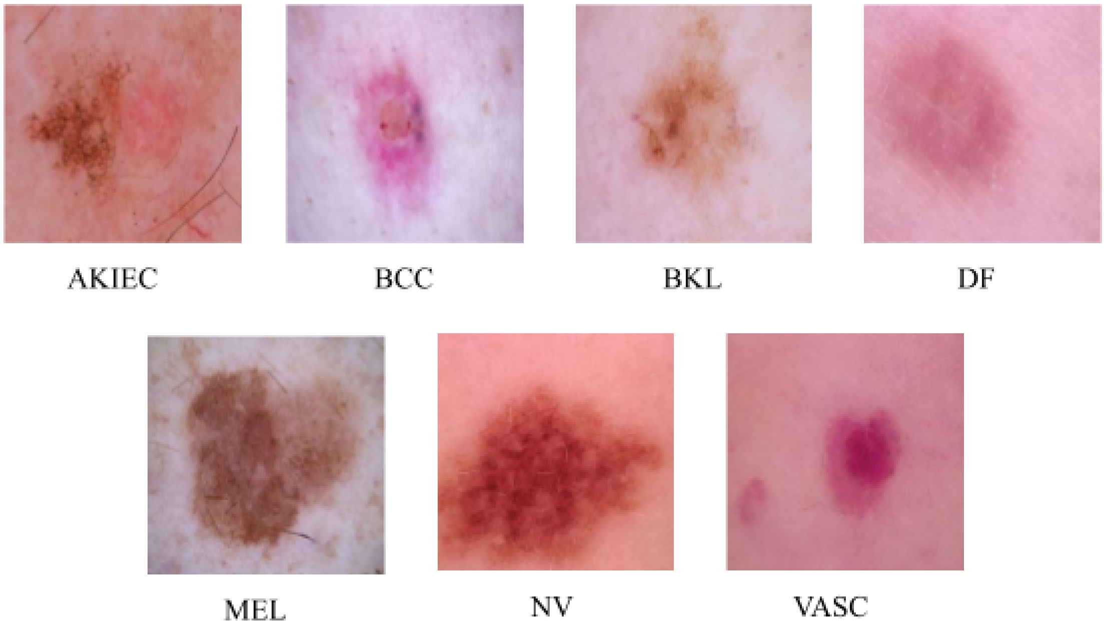
(a) JFT-300M vs ImageNet pre-training **(b) ViTs vs ResNets**

A comparison of ViT and CNN (including ResNets) models based on normalized layer number. ViT reach higher accuracies faster, requiring lesser computational resources (*AI-Scholar*).

Methods

Datasets

Several datasets are available online, but each has its own categories of skin disease. The HAM10000 dataset has seven types, including actinic keratoses/Bowen's disease (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (seborrheic keratoses and lichen-planus like keratosis) (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas, and hemorrhage) (VASC). This dataset is highly imbalanced, with melanocytic nevi making up 67% of the total images. The Stanford Dermatologically Diverse Images (DDI) dataset of non-fair images, in contrast, is relatively balanced, but the quality of some images are poor with visual distractions in the background. But all of the images in two datasets were verified by dermatologists, and their labels are accurate. To supplement this, the Fitzpatrick 17k dataset was used. Not all of its labels are verified, however, and only the labeled images were used. A new dataset was created out of all the three datasets mentioned above to minimize bias in the training data, pooling labels together as necessary to match the seven disease categories in the HAM10000 dataset. Several augmentation techniques were applied to this data before it was used to train AI models to compare accuracy and suitability.


AKIEC BCC BKL DF

MEL NV VASC

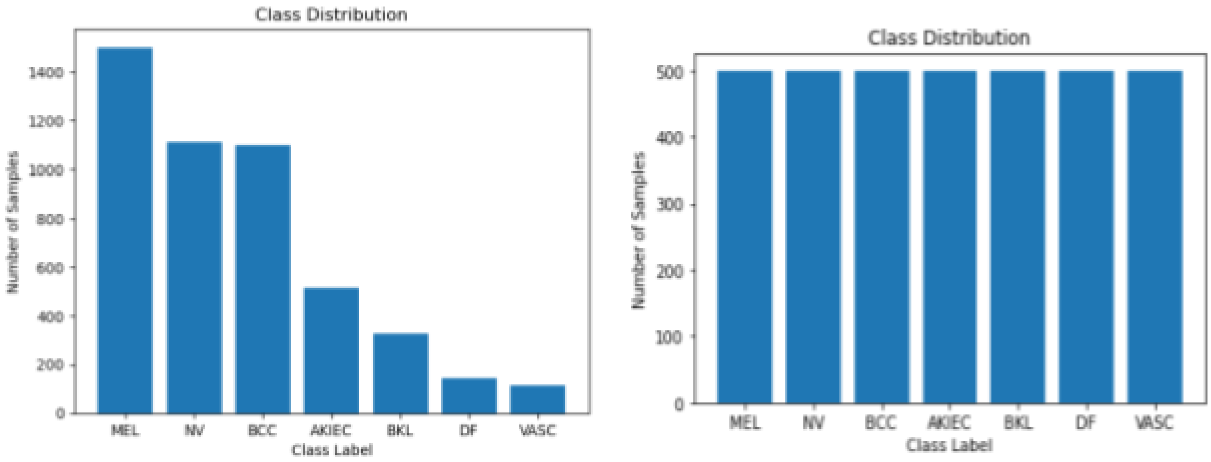
A sample of images in the HAM10000 dataset representing all skin condition types. (own image)

Data Preparation

Each of the three datasets had their own image sizes, resolutions, and color scales. To prevent the models from creating patterns on these distractions, data augmentation techniques were used. The dataset created out of these three source datasets was split into train, validation, and test using the 80:10:10 split. This allocated about 12,800 images for the train, 1,600 images for validation, and 1,600 images for test.

Dataset Creation

As the three datasets, HAM10000, DDI, and Fitzpatrick 17k, had their own directory structures and labeling systems, the first step was to find a standard. The HAM1000 dataset's labels were used as a basis for the other two. The images were organized into subdirectories by class within the dataset directory using the Python glob and shutil libraries. For both HAM10000 and DDI, the class information was given in a separate csv file containing the filenames of images and the disease labels. So, this file was used to match the filename of the image and find its corresponding class before putting those images into subdirectories by class. Then, the labels in the DDI had to be grouped in accordance with the HAM1000 labels. For example, the DDI dataset would have 'seborrheic-keratosis' and 'actinic-keratosis' under different labels, while HAM10000 had one class with the label 'AKIEC' which included both of these diseases. After grouping, the dataset was carefully monitored to ensure equal class distribution.



Class counts of non-fair DDI images before and after resampling. {0 : "MEL", 1 : "NV", 2 : "BCC", 3 : "AKIEC", 4 : "BKL", 5 : "DF", 6 : "VASC"} (own image)

Image Preprocessing

All the images in the newly created dataset, including training, validation, and testing subsets, were resized to 224 x 224 size to standardize the values. Then, random flipping, shearing, cropping, and rotation was applied to the training dataset to ensure that the models were not learning from a certain orientation, color, or other non trivial factors. However, an overuse of this method could create an excess of duplicates in the training set that could cause the model to overfit, learning to memorize the key features of images instead of learning to classify them. The pixel values of the RGB images were normalized by dividing all values by 255. This made it easier for the model to process the image information as an array of float values and lessened the processing time.

Training

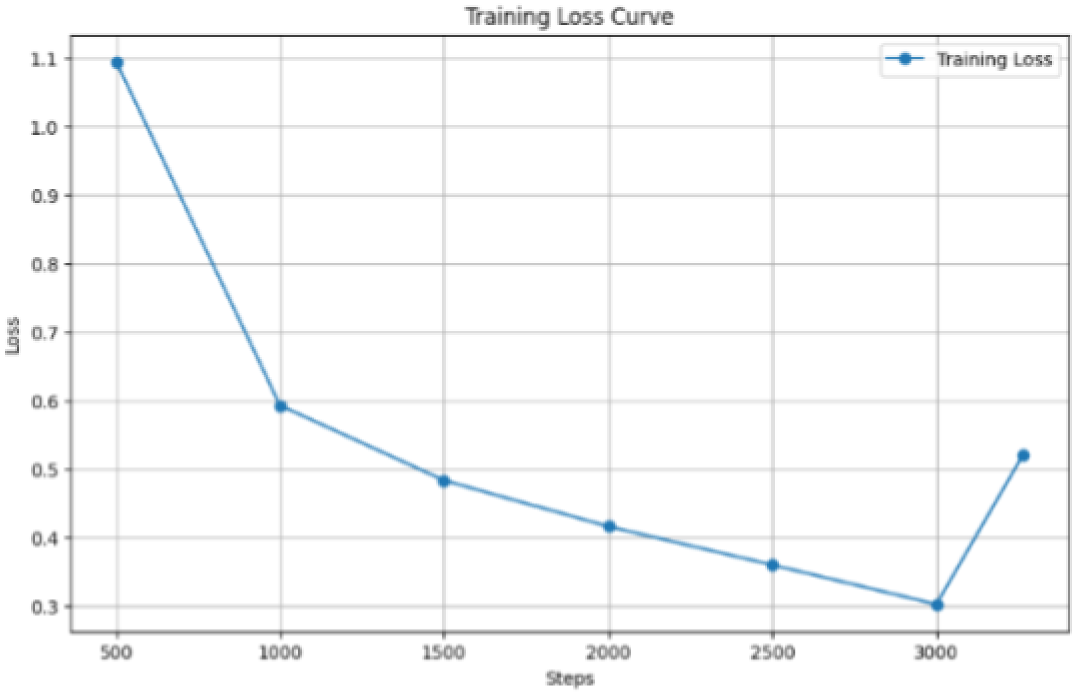
The data augmentation techniques were applied only to the training dataset, which was processed by a ViT model pre-trained on the ImageNet-21k dataset, a collection of 21,000 images. This highly effectivized the model, as it required only 5 epochs to train. The train loss curve was plotted to ensure the model was neither overfitting nor underfitting.

Hyperparameter Tuning

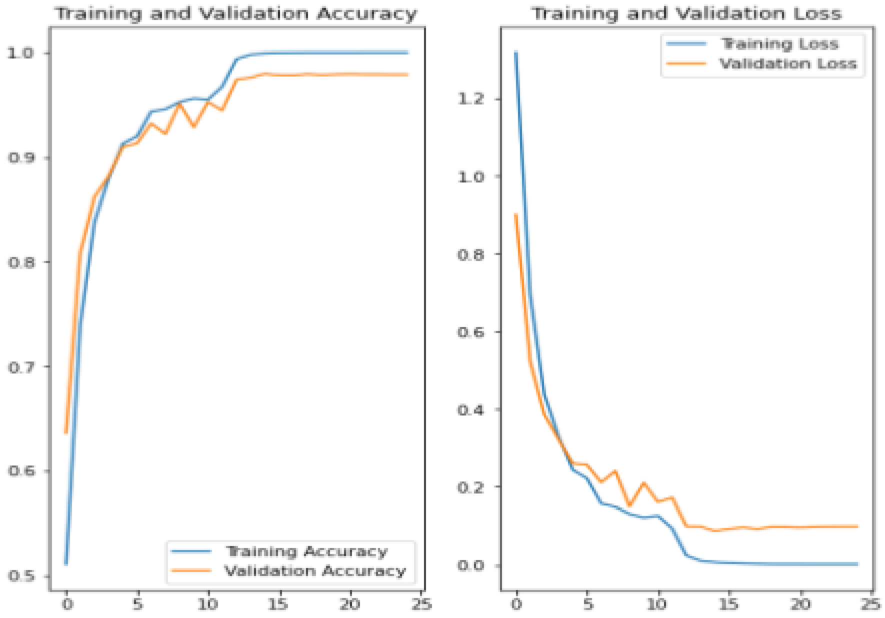
The learning rate, complexity, number of epochs, and optimizer type were modified based on the ViT's performance during training. The learning rate was set to a default of 2e-5, and lowered as needed based on the change in accuracy and loss values. The number of layers and hidden sizes per layer were initially set to those of the pretrained model, and were adjusted prior to training. The number of epochs was set to 5, and monitored by early stopping algorithms. For most of the trials, the ViT used all 5 epochs for optimal performance. The optimizer type was the default one from the pre-trained model. The hyperparameters were monitored using a Tensorboard, which plotted live changes in these values during training.

Model Architecture

The specific version of ViT used was the vit-base-patch16-224-in 21k model, which has many transformer blocks that have a self-attention layer and a feed-forward layer (Boesch). These layers process an image as a patch of images, making it computationally effective (Boesch). With 24 layers in total and a hidden size of 1024, the ViT requires an enormous amount of data if trained from scratch. However, the version used in this project was already pre-trained and fine-tuned on the 21,000 images in the ImageNet-21k dataset, enabling it to work even if the size of the dataset was 8,000, which is less than half of what is needed to optimize performance.



Training Loss Curve for Fair and No-fair Training Data (own image)



Train vs Validation Loss and Accuracy Average Curves for all models (own image)

Results

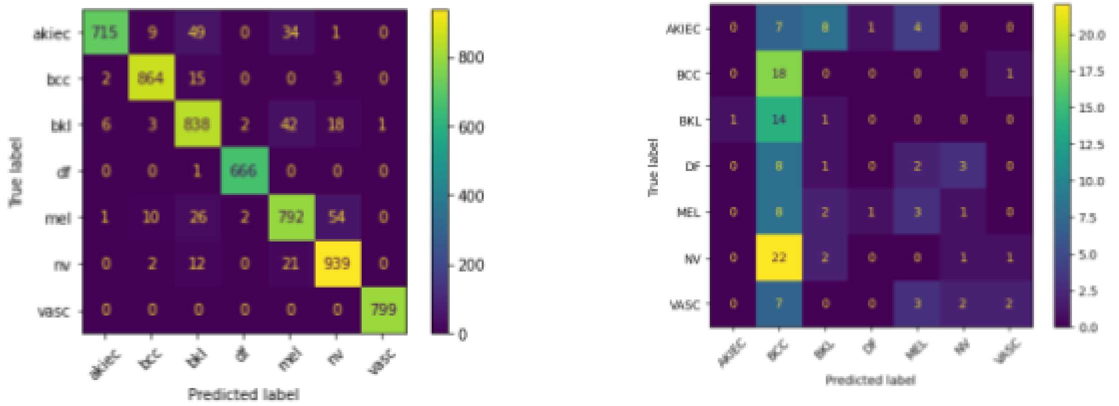
The ViT model obtained about 94.9% accuracy in classifying fair skin images when trained on only fair skin tones. The same model was tested on a real non-fair skin image dataset of 140 images, obtaining a baseline accuracy of 19.8%. Then, an equal amount of non-fair skin images, consisting of the 320 real non-fair skin images separate from the testing set and the simulated images obtained from the CycleGAN model, was added to the training data, doubling the size of the input for the ViT. After 5 epochs, the model achieved about 78.6% accuracy in classifying any skin tone and 52.1% accuracy in classifying real non-fair images, showing that the poor quality and quantity of the non-fair skin images were distracting the model.

Accuracy

The model did not show signs of overfitting or underfitting, as the validation and testing accuracy and loss curves were close to the training accuracy and loss curve. The values differed from each other by a maximum of 8%, showing consistency in the model. Precision and recall were calculated using the confusion matrix, which yielded about 0.68 and 0.56, respectively.

Confusion Matrix

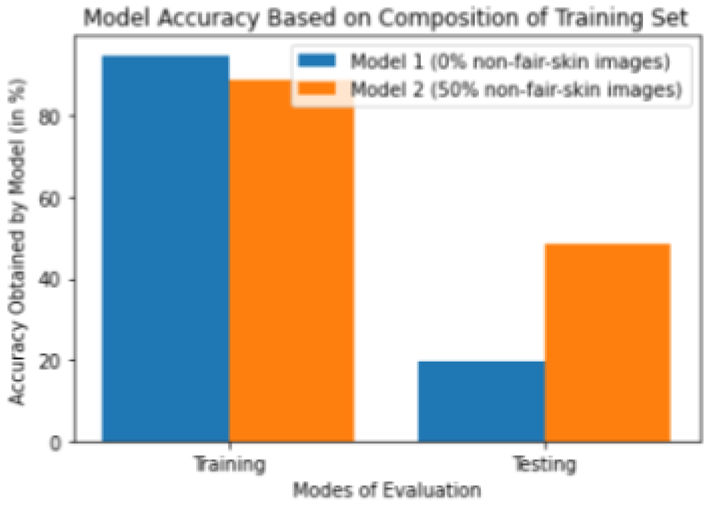
The confusion matrix was the easiest method to visually scan the performance of the ViT model. The perfect case, where the model classifies all images correctly, would have a 'staircase' pattern, as all the predicted labels would match the actual labels. As the accuracy decreases, this clear pattern gets more clouded.



	akiec	bcc	bkl	df	mel	nv	vasc
akiec	715	9	49	0	34	1	0
bcc	2	864	15	0	0	3	0
bkl	6	3	838	2	42	18	1
df	0	0	1	666	0	0	0
mel	1	10	26	2	792	54	0
nv	0	2	12	0	21	939	0
vasc	0	0	0	0	0	0	799

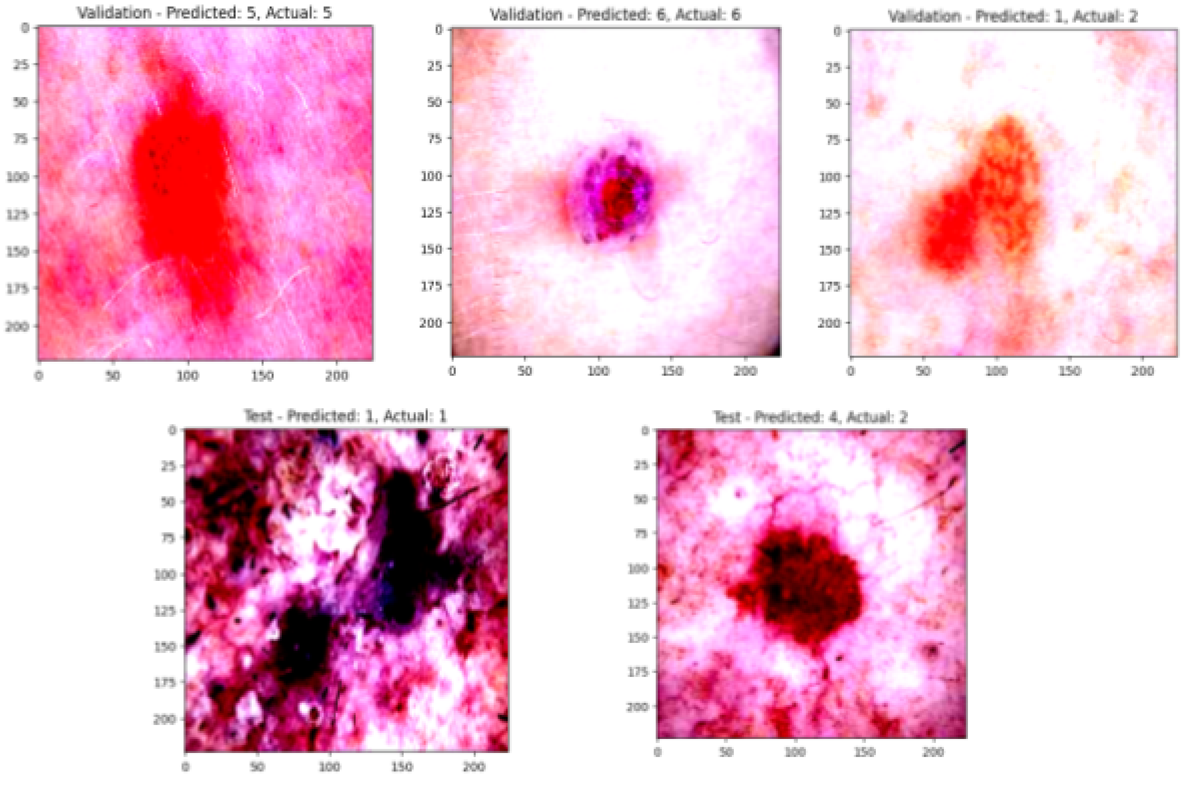
	AKIEC	BCC	BKL	DF	MEL	NV	VASC
AKIEC	0	7	8	1	4	0	0
BCC	0	18	0	0	0	0	1
BKL	1	14	1	0	0	0	0
DF	0	8	1	0	2	3	0
MEL	0	8	2	1	3	1	0
NV	0	22	2	0	0	1	1
VASC	0	7	0	0	3	2	2

Left: Confusion matrix of the ViT model trained and tested on fair images only.
Right: Confusion matrix of the ViT model trained and tested on both fair and non-fair images. (own image)



Model accuracy based on composition of training set (own image).

Sample Output


Validation - Predicted: 5, Actual: 5
Validation - Predicted: 6, Actual: 6
Validation - Predicted: 1, Actual: 2
Test - Predicted: 1, Actual: 1
Test - Predicted: 4, Actual: 2
0
25
50
75
100
125
150
175
200
0
50
100
150
200

Key{0 : "MEL", 1 : "NV", 2 : "BCC", 3 : "AKIEC", 4 : "BKL", 5 : "DF", 6 : "VASC"} (own images)

Discussion

Limitations

The main challenge of this research was to work with a limited number of non-fair skin images. All the data used in the project was derived from free publicly available resources, and sometimes, the quality had to be compromised. The Stanford Dermatologically Diverse Images (DDI) dataset, one of the rare datasets that contained exclusively non-fair images, often featured irrelevant objects surrounding the skin lesion, such as a table, wall, etc. that have the tendency to distract any image classification model. Several data augmentation techniques like cropping were used to ensure that the model was learning to predict based on the lesions themselves and not the background. In addition to these distractions, the zoom quality of some of the images was poor and this could have prevented the model from recognizing the lesion, lowering the accuracy of the ViT. Also, the CycleGAN generated images cannot substitute actual data, and there might have been patterns in the generated data (Khan et al.). While the results are pretty consistent with the estimated dermatologist accuracy in diagnosing skin disease, which ranges from 48 to 77%, the findings of this research should be used with caution, as AI models cannot be guaranteed to classify correctly for all cases.

Conclusion

The lack of representation of non-fair skin tones in the datasets used to train AI models that assist dermatologists is alarming (De et al.). This can lead to severe consequences for minorities: failed diagnosis, late detection, costly procedures, and slimmer chances of survival. One of the direct ways to combat the issue of data scarcity is to add more samples of the necessary data (Yan et al.). Perhaps, crowdfunding could be used to create more non-fair skin datasets. Certified dermatologists could annotate these images, ensuring quality data is available for training. In the event of scarcity, to bridge the gap between the accuracy of AI skin disease diagnosis by high-performing Vision Transformer models between various skin tones, I utilized another AI model, CycleGAN to generate simulated non-fair images to supplement the scarce real non-fair skin data available online. The overall accuracy of 68.4% across all skin tones is comparable to that of dermatologists, demonstrating that this method can potentially be used to balance fair and non-fair skin images. Beyond skin care, this data augmentation technique can be used in any industry that faces the problems of a majority and minority class.

Selected References

Balch, B. (2022, July 21). *Why are so many black patients dying of skin cancer?*. AAMC. https://www.aamc.org/news/why-are-so-many-black-patients-dying-skin-cancer. Accessed March 25 2023.

Boesch, G. (2023, March 7). *Vision transformers (VIT) in image recognition - 2023 guide*. viso.ai. https://viso.ai/deep-learning/vision-transformer-vit/. Accessed September 9 2023.

Burden of skin disease. American Academy of Dermatology. (n.d.). https://www.aad.org/member/clinical-quality/clinical-care/bsd. Accessed June 15 2023.

De, A., Sarda, A., Gupta, S., & Das, S. (2020). *Use of artificial intelligence in dermatology*. Indian Journal of Dermatology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7640800/. Accessed October 29 2023.

Historoid. (2023, November 8). *Why are vision transformers so high performance?*. AI. https://ai-scholar.tech/en/articles/transformer/transformer-vs-cnn. Accessed March 25 2023.

Khan, M. H.-M., Sahib-Kaudeer, N. G., Dayalen, M., Mahomedaly, F., Sinha, G. R., Nagwanshi, K. K., & Taylor, A. (2022). (publication). *Multi-Class Skin Problem Classification Using Deep Generative Adversarial Network (DGAN)*. Retrieved July 25, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8995545/pdf/CIN2022-1797471.pdf. Accessed July 22 2023.

Yan, Y., Hong, S., Zhang, W., & Li, H. (2022). Artificial Intelligence in skin diseases: Fulfilling its potentials to meet the real needs in dermatology practice. *Health Data Science*, *2022*. https://doi.org/10.34133/2022/9791467. Accessed September 15 2023.

Yotsu, R. R., Ding, Z., Hamm, J., & Blanton, R. E. (n.d.). *Deep learning for AI-based diagnosis of skin-related neglected tropical diseases: A pilot study*. PLOS Neglected Tropical Diseases. https://journals.plos.org/plosntds/article?id=10.1371%2Fjournal.pntd.0011230. Accessed October 29 2023.