
Analyzing Skin Tone Bias in Deep Neural Networks for Skin Condition Diagnosis

SANTHOSH GADIPELLY, MS, Data Science.
RAVI CHANDRA MADAMANCHI, MS, Data Science.
Florida International University, Miami, FL.

ABSTRACT

Millions of people worldwide are impacted by a variety of skin issues that can negatively impact their health and quality of life. It is essential to make an accurate diagnosis of skin conditions in order to provide efficient care and enhance patient outcomes. However, because darker skin tones are underrepresented in dermatology, it can be challenging to accurately diagnose skin issues, particularly for people with darker skin tones.

The creation of deep learning models for skin diagnosis has the potential to revolutionize the field of dermatology by offering accurate, automated, and simple diagnoses for a range of skin disorders. By evaluating how well these models perform across a range of skin tones, we can identify any potential biases or limitations and work to develop more inclusive and accurate models.

Fitzpatrick17k dataset study on deep learning approaches for skin disease diagnosis is important since it offers insights into how well these models function with various skin kinds and colors. The results of this study can assist construct deeper learning skin diagnosis models that are more precise and inclusive, which can enhance patient outcomes and lessen healthcare inequities.

The significance of this discovery rests primarily on its potential to increase the precision and accessibility of skin diagnostics for persons with all shades and varieties of skin. We can enhance patient outcomes, lessen healthcare disparities, and advance health equity by offering accurate and accessible diagnoses for a variety of skin disorders.

INTRODUCTION

By delivering precise, automated, and easily accessible diagnoses for a variety of skin disorders, the development of deep learning models for skin diagnosis has the potential to transform dermatology. Using the Fitzpatrick17k dataset, we compare the performance of VGG16, DenseNet121, and InceptionNet, three deep neural networks, for identifying skin disorders.

We investigate various strategies, including data augmentation, altering contrast, transfer learning, and fine-tuning, to overcome potential biases in the performance of these models across distinct skin types. The goal of the study is to determine how much these methods can enhance the models' ability to diagnose skin problems in people with various skin tones.

Using accuracy and skin tone, we compare the three models' performance. We evaluate the study's findings and talk about how they could be used to build deeper learning skin diagnosis models that are both more accurate and inclusive.

The results of this experiment can serve as a roadmap for the creation of deeper learning models for skin diagnosis that are more inclusive and accurate, which can enhance patient outcomes and lessen healthcare inequities. This experiment demonstrates the possibility of several techniques to overcome potential biases and enhance the performance of these models, emphasizing the significance of considering the diversity of skin types and colors when constructing deep learning algorithms for detecting skin disorders.

MOTIVATION AND PROJECT APPLICATIONS

Due to implicit biases, a lack of diversity in research and training data, or even both, people with darker skin tones may receive delayed or incorrect diagnoses. People from racial and ethnic minorities have lower rates of skin disorder diagnosis and treatment, as is well known, and there are healthcare disparities in these areas.

The creation of deep learning models for skin diagnosis has the potential to transform dermatology by providing accurate, automated, and simple diagnoses for a range of skin disorders. However, in order to ensure that these models are precise and inclusive, it is crucial to evaluate their performance across a variety of skin tones and types.

In this study, we use the Fitzpatrick17k dataset to assess how well deep learning models do in diagnosing skin problems in individuals with various skin tones. We can explore methods for creating more precise and inclusive deep-learning models for skin diagnostics by evaluating the efficacy of various tactics to address potential biases and enhance the performance of these models.

The results of this experiment can serve as a roadmap for the creation of deeper learning models for skin diagnosis that are more inclusive and accurate, which can enhance patient outcomes and lessen healthcare inequities. We can improve access to high-quality care for people with skin problems, regardless of their skin type or color, by addressing potential prejudices and creating more precise and inclusive models.

OBJECTIVES AND DELIVERABLES

There are three main objectives for this project,

1. Develop deep neural network models that were honed using photographs of people with varied skin tones and assess their efficacy using images of people with diverse skin tones.

Deliverables:

- a. The Fitzpatrick17k dataset was preprocessed such that it could be used for model training and testing.
- b. Putting into practice three deep learning models—VGG16, DenseNet121, and InceptionNet—that were trained on the preprocessed dataset.

c. Testing the models on pictures of people with various skin tones to determine how accurate each model is on various skin tones.

d. Code used to implement the models and evaluate their accuracy must be documented.

2. Determine whether there are any inconsistencies between different skin tones and evaluate the accuracy results.

Deliverables:

A comparison of the accuracy scores for each model on the range of skin tones.

b. Finding any inaccuracies in accuracy across various skin tones.

c. Findings and recommendations about the accuracy discrepancies and the causes of them.

d. Recording the analyses and conclusions drawn from evaluating the accuracy outcomes.

3. Explore several strategies to combat bias in model results and lessen accuracy differences.

Deliverables:

a. Investigating different strategies, such as data augmentation, transfer learning, and fine-tuning, to lessen accuracy differences and combat biased model results.

b. Implementing the selected methods to update the deep neural network models and reassess their efficacy on photos with various skin tones.

c. A comparison of the accuracy results acquired before and after applying the procedures to gauge how well they work in minimizing discrepancies and biased findings.

d. Knowledge gained and conclusions drawn regarding the efficacy of the methods in minimizing accuracy discrepancies and biased results.

e. The code used to implement the techniques and the accuracy results attained both before and after their use are documented.

TOOL USED AND DATA COLLECTED

TOOLS AND SYSTEM

Python 3.0 was used for this project in a Jupyter Notebook environment on two different laptops: an HP ProBook 440 G6 with an Intel i5-8th generation CPU running at 1.6 GHz, an Apple M1 chip, eight cores, eight gigabytes of RAM, and a 512 TB SSD; and a MacBook Air. We used Draw.io to make architectural diagrams for the models. Preprocessing, exploratory data analysis, constructing models, and testing all used the following Python libraries:

- pandas
- numpy
- os
- matplotlib.pyplot
- seaborn
- tensorflow
- tensorflow.keras.preprocessing.image
- tensorflow.keras.applications.inception_v3
- tensorflow.keras.layers
- tensorflow.keras.models
- sklearn.model_selection
- keras.preprocessing.image
- keras.models
- keras.layers
- keras.applications.densenet

Additionally, Draw.io, Microsoft Excel, PowerPoint, and Word were used to create visualizations, presentations, and documentation to effectively communicate the results of the project. These tools facilitated the organization and presentation of data, analysis, and insights in a clear and visually engaging manner.

DATASET

The Fitzpatrick17k dataset is an extensive collection of clinical photos that can be used for a variety of academic projects as well as computer vision and dermatology applications. The collection gathers 16,577 clinical photos with captions for skin conditions and skin types based on the Fitzpatrick grading system from two free open-source dermatology atlases, DermaAmin and Atlas Dermatologico.

Board-certified dermatologists evaluated the dataset's quality and determined how well the photos served as diagnostics. The dataset was more dependable for research purposes as a result of this evaluation method, which helped to detect and confirm the error rate in the dataset.

The Fitzpatrick17k dataset's developers focused on the most prevalent dermatological disorders found in the two source atlases when choosing photos for annotation. Due to factors like low image quality, an overly wide classification, or the depiction of rare genodermatoses, they omitted some groups. As a result, there are 114 conditions in the final dataset, each comprising at least 53 photos.

Fitzpatrick skin type classifications were applied to the photos by human annotators from Scale AI throughout the annotation process. The Fitzpatrick scale has some limitations in terms of capturing the complete range of skin types, despite being widely used for categorizing sun sensitivity and tailoring clinical medication according to skin phenotype. When evaluating algorithmic fairness using the dataset, this should be kept in mind.

The imbalance in skin type representation, with more photos showing lighter skin types than darker ones, is a significant feature of the Fitzpatrick17k dataset. The distribution of skin types among labels for skin conditions also shows this disparity.

Despite these drawbacks, the Fitzpatrick17k dataset is a useful tool for researchers and programmers working on computer vision, dermatology, and machine learning models. The dataset contributes to a better understanding of different skin types and disorders, which will eventually result in more precise and inclusive technology.

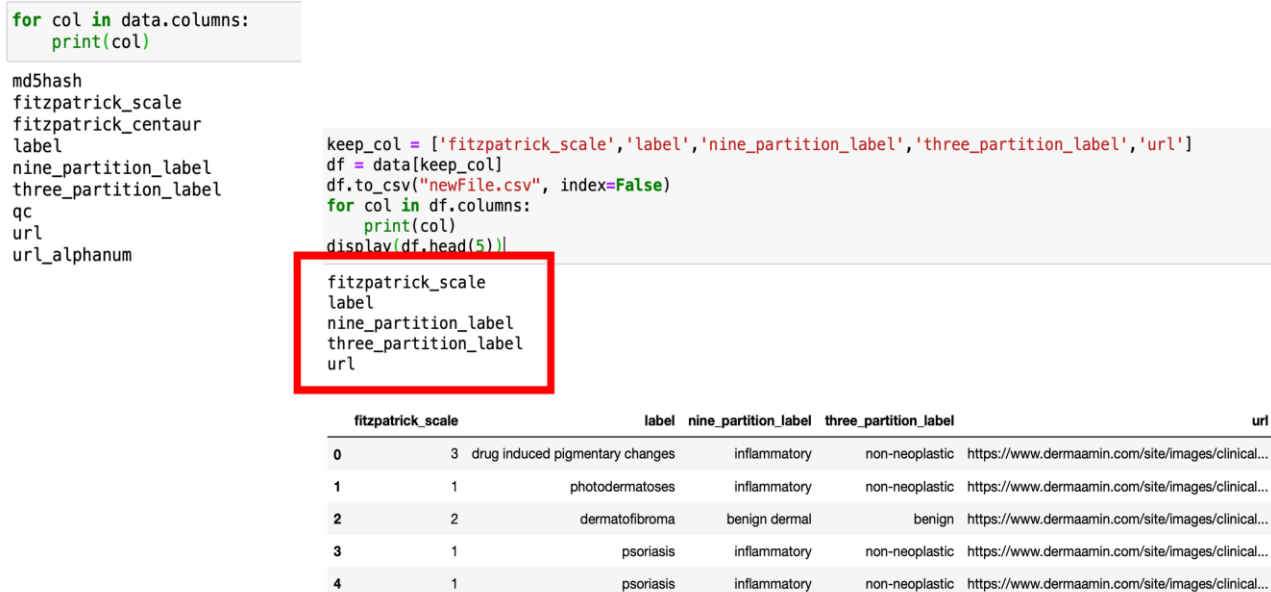


Figure 1

According to Figure 1, the Fitzpatrick17k dataset initially had 9 columns, including "mdhash," "Fitzpatrick_scale," "Fitzpatrick_centaur," "Label," "nine_partition_label," "Three_partition_label," "qc," "url," and "url_alphanum." However, during the preprocessing stage, some columns were deemed unnecessary and removed, leaving only the columns required for the project. These columns are "Fitzpatrick_scale," "Label," "Nine_partition_label," "Three_partition_label," and "url."

- i. **Fitzpatrick_scale:** This column has values ranging from 1 to 6 based on the Fitzpatrick scale.

The Fitzpatrick scale was developed in 1975 by Thomas B. Fitzpatrick, MD, PhD, as a numerical classification system to predict how different skin types will respond to ultraviolet (UV) light. The scale was initially developed for use in dermatology clinical practice and research, but it has since been used in fields like genetics, computer vision, and the manufacture of cosmetics.

The Fitzpatrick scale divides human skin into six groups, ranging from very fair to very dark, based on how it responds to exposure to sunlight. These classifications are influenced by both hereditary and environmental variables, including the quantity of melanin in the skin and exposure to sunshine. The six groups are as follows:

Type I (very fair): Never tans and always burns. Freckles, red or blonde hair, and pale white skin are typical skin characteristics.

Type II (fair): Typically burns, barely tans. Typically, people have white skin, light hair, and blue or green eyes.

Type III (medium): Occasionally burns, uniformly tans. Any eye or hair color goes well with cream-white skin.

Olive: Type IV (rarely burns, tans well). With dark hair and eyes, mild brown skin is the norm.

Type V (brown): Tans easily and very rarely burns. Typically, people have dark brown skin, brown eyes, and black hair.

Dark brown or black Type VI: Never burns, always tans quickly. With brown or black eyes and black hair, very dark brown to black skin is the norm.

The Fitzpatrick scale can be used to predict the likelihood of developing skin cancer as well as to determine the best course of treatment for various skin issues. It is also used in the manufacturing of cosmetic products to guarantee that goods are suitable for a variety of skin types and colors. The Fitzpatrick scale has also developed into a crucial tool in computer vision, where it is employed to assess the fairness and correctness of model algorithms for various skin tones.


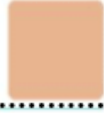
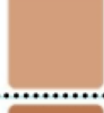
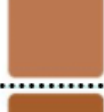


SKIN TYPE		SKIN COLOR	REACTION TO SUN	
			SUNBURN	TANNING
I		Light, pale white	Always burns	Never Tans
II		White, fair	Usually burns	Tans with difficulty
III		Medium, white to olive	Sometimes mild burns	Gradually tans to olive
IV		Beige olive, moderate brown	Rarely burns	Easy tan to moderate brown
V		Brown, dark brown	Very rarely burns	Tans very easily
VI		Very dark brown to black	Never burns	Always tans

Figure 2 : Fitzpatrick scale

The following figure shows the data distribution with respect to Fitzpatrick_scale.

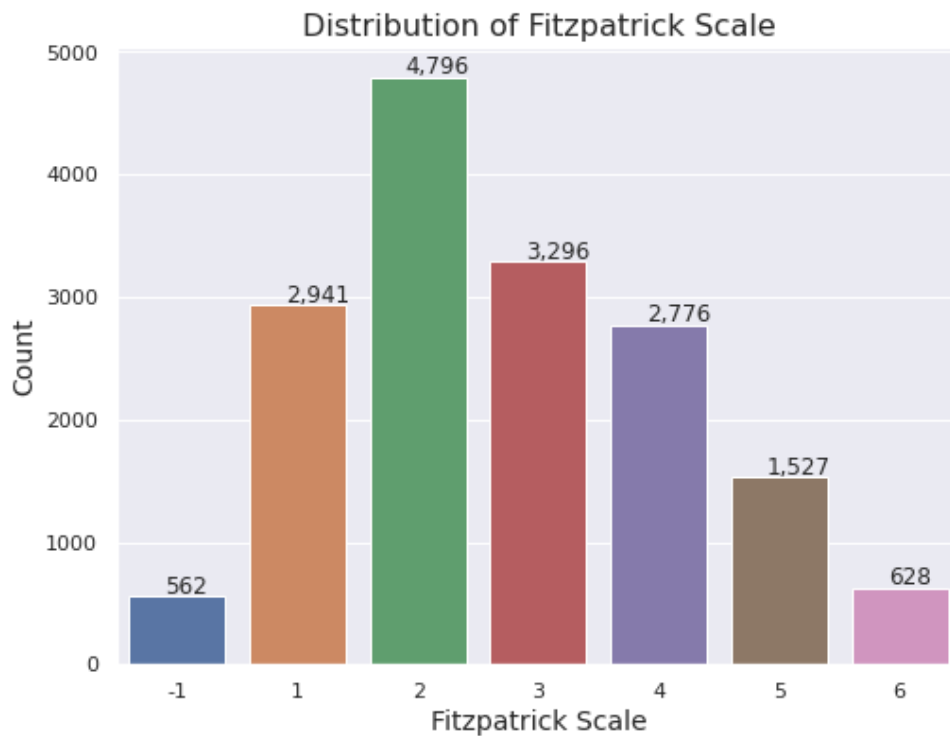


Figure 4

ii. **Label :**

The "Label" column in the Fitzpatrick17k dataset contains the specific skin condition that has been diagnosed in the corresponding image. There are 114 unique values for the "Label" column, representing the various skin conditions that were selected for inclusion in the dataset. The inclusion of these specific labels makes the dataset valuable for research and development of machine learning models and applications related to dermatology and skin health.

iii. **Three_partition_label :**

Non-neoplastic, Benign, and Malignant are the three distinct values of the three_partition_label variable. Based on the photo's common traits, such as the presence of tumors or other abnormalities, these labels can be used to group the images into several categories. This category is a helpful tool for dermatological research and clinical practice since it can offer insightful information about the overall occurrence and distribution of skin disorders across several categories.

Non-neoplastic, benign, and malignant are three general categories used to classify skin lesions based on their characteristics and potential risk for cancerous growth.

Non-neoplastic Growths known as lesions are unrelated to cancer or precancerous changes. These lesions cover a variety of skin conditions like rosacea, eczema, and psoriasis in

addition to skin wounds like burns or cuts. Numerous medications and therapies can be used to treat non-cancerous lesions, and they don't significantly raise the risk of getting cancer.

Benign lesions are those that, albeit not malignant, could spread or become a problem if untreated. Different kinds of moles, cysts, and skin tags are among these lesions. They are not thought to be malignant, but if they enlarge or show symptoms, they might need medical care.

Growths known as **malignant lesions** either have the potential to or are actively undergoing the development of cancer. These lesions include many skin cancers, including melanoma, basal cell carcinoma, and squamous cell carcinoma. Malignant lesions must be treated right away because, if not, they could be fatal.

Making the distinction between benign and malignant tumors is crucial for accurate diagnosis and treatment. While benign and malignant lesions need to be closely monitored and treated to prevent turning into cancer or other serious health issues, non-neoplastic lesions may not pose a significant risk to health.

The following figure shows the data distribution with respect to three_partition_label.

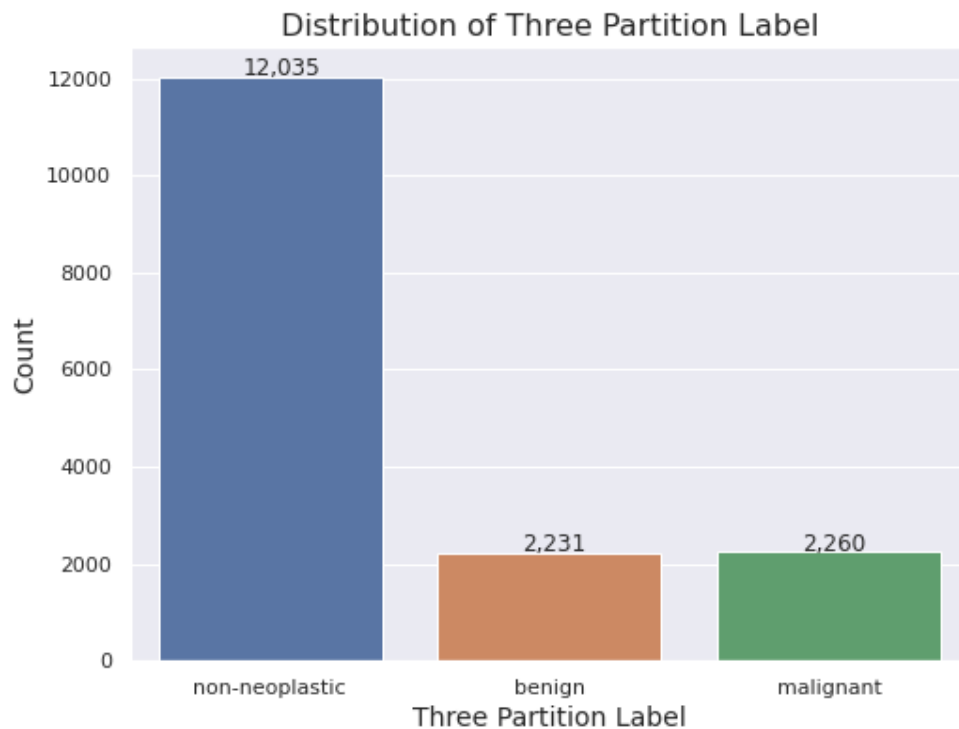


Figure 3

iv. **Nine_partition_label :**

In comparison to the "Three_partition_label" column, the Fitzpatrick17k dataset's "Nine_partition_label" column offers a more thorough classification of skin disorders. There

are nine possible values in this column, each of which represents a different class of skin lesions. These groups include:

Inflammatory: Dermatitis, psoriasis, and lupus are just a few of the inflammatory skin illnesses included in this group.

Malignant epidermal: Squamous cell carcinoma is one of the skin cancers that fall under this classification and develop in the epidermis.

Genodermatoses: This group covers some inherited skin disorders including epidermolysis bullosa and xeroderma pigmentosum.

Benign dermal: These non-cancerous skin lesions include hemangiomas and lipomas, which develop from the skin's deeper layer.

Benign epidermal: This group encompasses a variety of non-cancerous skin lesions that develop from the epidermis, including verrucae and seborrheic keratosis.

Malignant melanoma: This group contains numerous skin malignancies, including melanoma, that develop in the skin's pigment-producing cells.

Benign melanocyte: These non-cancerous skin lesions, which include nevi and lentigines, develop from the skin's pigment-producing cells.

Malignant cutaneous lymphoma: This group contains numerous skin malignancies, including mycosis fungoides, that develop from the skin's white blood cells.

Malignant dermal: This group comprises several skin tumors that develop in the skin's deeper layers, like dermatofibrosarcoma protuberans.

For a more detailed classification of skin lesions and to get an additional understanding of the diagnosis and management of skin disorders, the "Nine_partition_label" column can be helpful.

v. URL :

In the URL column, we have a link to the images for that record. We can use the link to download the image.

PREPROCESSING OF DATA

We used four main procedures in the preprocessing phase of our project to get the Fitzpatrick17k dataset ready for more analysis.

First, we used the links in the "url" column to download each and every image in the collection. We saved all 16,577 pictures to a "fitzpatrick_images" folder on our computer. To get access to the photos for later processing and analysis, this step is essential.

Second, we added a new column called "local_filename" with the local path of each image to its related record. This approach made it possible for us to maintain track of each image's position on our local computer and to quickly access the photos for use in further processing.

Thirdly, we discovered a few URLs that were inaccessible or unresponsive for a predetermined period of time. Records whose photos, for any reason, could not be downloaded were erased. In order to have a complete dataset with all of the photos intact for subsequent analysis, it is imperative that we take this step.

Finally, we noticed that some data had a Fitzpatrick scale value of -1. These images are represented by records whose skin tones were indistinguishable for a variety of reasons. These records were taken out of the dataset because we could not utilize them to conduct our analysis.

We were left with a final dataset of 15,956 entries once all the preprocessing procedures were finished, each of which had an associated image and a local filename. These records were prepared for additional examination utilizing several machine-learning approaches and algorithms.

FURTHER DATA CLEANING

We discovered some discrepancies in the Fitzpatrick17k dataset after preprocessing it, necessitating additional cleaning. We performed various data-cleaning procedures to make sure that our dataset was error-free and prepared for future research.

The first step was to ensure that the dataset's photos all had proper filenames. Special characters in filenames can occasionally make it challenging to access the photos, which could lead to issues during analysis. Therefore, in order to ensure that all photos could be easily found, we eliminated any records with incorrect filenames.

In the following phase, it was determined whether any empty or non-image files had been downloaded. Such issues can occasionally happen when downloading a big number of files. To make sure that only accurate photos were included in our final dataset, we deleted any entries with such problems.

In the third stage, corrupt image detection was performed. To accomplish this, we attempted to open each image using the "image.open()" function. Records with damaged photos were discarded since they couldn't be used for analysis.

The dataset was examined for duplicate photos in the fourth stage. The dataset would become redundant if several records had the same image. To make our dataset more efficient for analysis, we eliminated any duplicate photos.

Checking each image's file extensions was the final and fifth stage. We made sure that every picture in the dataset has a legitimate extension, like .png, .jpg, or.jpeg. Records containing photos with incorrect extensions were eliminated since the deep neural network model couldn't access these images.

We ended up with a final dataset of 12,215 records that was clean and prepared for additional analysis using deep neural network models after carrying out all these data cleaning stages. These procedures assisted us in ensuring the dataset's accuracy and dependability, which are essential for any machine-learning research.

IMPLEMENTATION

MODEL 1: VGG16

The goal here is to create a deep learning model based on the VGG16 architecture to categorize different skin types according to the Fitzpatrick scale. The model will be tested for accuracy in predicting skin types after being trained on a unique dataset.

Dataset: The dataset 'final_dataset.csv' which is the updated dataset after all the preprocessing and data cleaning, used in this project includes labels for each image file path that reflect different skin types according to the Fitzpatrick scale. To guarantee equitable representation of all classes, the dataset is divided into training and validation subsets using a stratified technique.

The photos are downsized to 224x224 pixels, and the pixel values are rescaled to the [0, 1] range as part of the data preprocessing and augmentation procedure. To increase the model's ability to generalize to new data, data augmentation techniques such as horizontal flipping, rotation, shear, and zoom are applied to the training dataset.

Model Construction: The model is based on the VGG16 architecture, a deep-learning model for image classification that has already been trained. The following layers make up the model:

(include_top=False, weights='imagenet,' input_shape=(224, 224, 3), VGG16 base model)

Dense layer with 1024 neurons and activated ReLU in the flatten layer

Dense output layer with 114 neurons (corresponding to the number of classes), a dropout layer with a rate of 0.5, and Softmax activation.

Model Education:

The Adam optimizer and categorical cross-entropy loss are used in the model's construction, and its learning rate is 1e-5. Using the training dataset and validation dataset, it is trained over 20 iterations.

Model Evaluation: Using the validation dataset, the model's effectiveness is assessed. Calculations and reports are made about test accuracy overall. Additionally, the model's forecasts for the validation dataset, along with the matching Fitzpatrick scale values and ground truth labels, are recorded in a new DataFrame. The CSV file 'predicted_labels.csv' is created from this DataFrame for further study.

The below figure shows you the architecture of the VGG16 model we used.

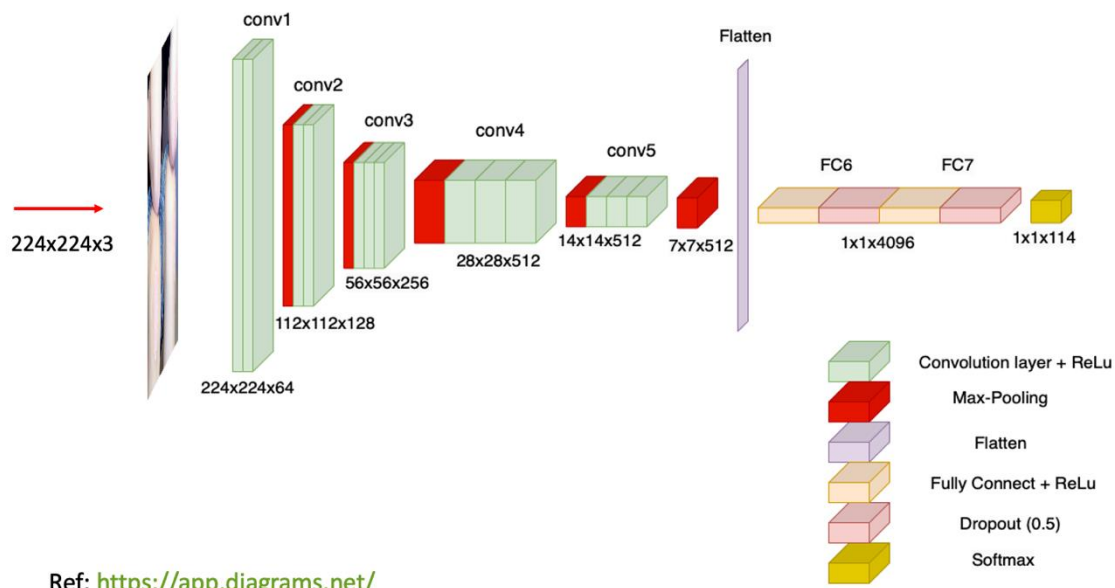


Figure 4 : VGG16 ARCHITECTURE

MODEL 2: DenseNet

The purpose here is to create a deep learning model based on the DenseNet121 architecture that categorizes skin types according to the Fitzpatrick scale. The algorithm will be tested for its accuracy in predicting skin types after being trained on a unique dataset.

Dataset: The dataset 'final_dataset.csv' which is the updated dataset after all the preprocessing and data cleaning, used in this project includes labels for each image file path that reflect different skin types according to the Fitzpatrick scale. To guarantee equitable representation of all classes, the dataset is divided into training and validation subsets using a stratified technique.

The photos are downsized to 224x224 pixels, and the pixel values are rescaled to the [0, 1] range as part of the data preprocessing and augmentation procedure. With a 20% validation set, the training dataset is further divided into training and validation subsets. Techniques for enhancing data are not used in this instance.

Model Construction: The model is based on the DenseNet121 architecture, a deep-learning model for image classification that has already been trained. The following layers make up the model:

(Include_top=False, pooling='avg', weights='imagenet') DenseNet121 basic model

Dense output layer with Softmax activation and 114 neurons (equivalent to the number of classes).

Model Education:

The model is built using categorical cross-entropy loss and the Adam optimizer, with the DenseNet121 layers frozen to use the pre-trained weights. Using the training and validation datasets, the model is trained over a period of 20 iterations.

Model Evaluation: The model's effectiveness is assessed using the test dataset. Calculations and reports are made about test accuracy overall.

MODEL 2: InceptionNet

InceptionV3 architecture will be used in this project to create a deep-learning model that categorizes skin types based on the Fitzpatrick scale. The algorithm will be tested for its accuracy in predicting skin types after being trained on a unique dataset.

Dataset: The dataset 'final_dataset.csv' which is the updated dataset after all the preprocessing and data cleaning, used in this project includes labels for each image file path that reflect different skin types according to the Fitzpatrick scale. To guarantee equitable representation of all classes, the dataset is divided into training and validation subsets using a stratified technique.

The photos are downsized to 224x224 pixels, and the pixel values are rescaled to the [0, 1] range as part of the data preprocessing and augmentation procedure. The algorithm contains a stage for optional data augmentation that applies shear, zoom, and horizontal flip transformations.

Model construction: The model is based on the InceptionV3 architecture, a deep-learning model for image categorization that has already been trained. The following layers make up the model:

(Include_top=False, weights='imagenet') InceptionV3 base model

Layer GlobalAveragePooling2D

1024 neurons in a dense layer, exhibiting ReLU activity.

dense output layer with Softmax activation and several neurons equal to the number of classes

Model Education:

The model is built using categorical cross-entropy loss and the Adam optimizer, with the InceptionV3 layers frozen to use the pre-trained weights. The training dataset is used to train the model for 20 epochs, and the testing dataset is used to validate it.

Model Evaluation: The model's effectiveness is assessed using the test dataset. Calculations and reports are made about test accuracy overall.

RESULTS

For each of the three models, we employed four distinct holdout sets to assess how well our deep neural network models performed. In figure 5, we evaluated the VGG16 model's performance on the four holdout sets.

Figure 5 presents the findings of our analysis and gives a summary of the model's performance on each of the four holdout sets. These findings made it easier for us to comprehend how well the VGG16 model performed using the Fitzpatrick17k dataset to forecast skin disorders.

We were able to assess the performance of our model under various circumstances and confirm that the model's performance was constant across several test sets by employing multiple holdout sets. This method gave us a more thorough grasp of the model's performance and gave us important information about its advantages and disadvantages.

Holdout set	Random	Fst 3-6	Fst 1,2,5,6	Fst 1-4
# Train images	9772	6913	3984	1318
# Test images	2443	5302	8231	10897
Overall	26.3%	15.8%	15.6%	9.8%
Type 1	29.2%	-	11.8%	5.7%
Type 2	38.5%	-	14.0%	6.4%
Type 3	13.8%	21.3%	-	10.8%
Type 4	12.3%	18.4%	-	12.6%
Type 5	8.6%	9.3%	15.9%	-
Type 6	33.8%	6.1%	11.6%	-

Figure 5 : VGG16 Results

The random holdout set is a randomly sampled set of images. The 3 Fitzpatrick holdout sets are selected according to Fitzpatrick labels. In all cases, the training data are the remaining non-held-out images from the Fitzpatrick 17k dataset.

Interesting conclusions were gained from our research of the Fitzpatrick17k dataset performance of the VGG16 model. In the beginning, we assessed the model's overall accuracy using training and testing datasets without any restrictions. The model's total accuracy was relatively low, coming in at just 26.3%.

A random pattern of accuracy ratings was discovered after further study of the holdout sets for various Fitzpatrick skin types. Nevertheless, we noticed a recurring pattern in the accuracy ratings for various Fitzpatrick skin types. The model's accuracy was highest for FST 3 photos while analyzing the FST 3-6 holdout set, and it gradually declined as we progressed from FST 3 to FST 6.

The FST 1, 2, 5, 6, and FST 1-4 holdout sets showed a similar pattern, as well. This data shows that the Fitzpatrick skin types that are most similar to the skin types utilized in training had the highest model accuracy. This result emphasizes the DNN models' predilection towards particular skin tones.

In conclusion, we discovered that the VGG16 model performed better for skin types close to those used in training while having low overall accuracy. These results show that when DNN models are trained on medical image datasets, biases in the models must be addressed.

Holdout set	Random	Fst 3-6	Fst 1,2,5,6	Fst 1-4
# Train images	9772	6913	3984	1318
# Test images	2443	5302	8231	10897
Overall	23.8%	16.6%	16.9%	17.3%
Type 1	27.4%	-	14.6%	16%
Type 2	23.9%	-	16.5%	16.8%
Type 3	22.9%	12.1%	-	17.3%
Type 4	22.5%	11.2%	-	17.8%
Type 5	18.5%	10%	15.1%	-
Type 6	19.3%	5.6%	7.8%	-

Figure 6 : DenseNet Results

Holdout set	Random	Fst 3-6	Fst 1,2,5,6	Fst 1-4
# Train images	9772	6913	3984	1318
# Test images	2443	5302	8231	10897
Overall	20.3%	13%	13.7%	13.87%
Type 1	19.1%	-	14.8%	12.0%
Type 2	23.3%	-	16.4%	12.1%
Type 3	20.7%	11%	-	15.42%
Type 4	16.1%	11.9%	-	14.23%
Type 5	16.5%	5.8%	14.9%	-
Type 6	26.9%	10.2%	5.7%	-

Figure 7 : InceptionNet Results

The accuracy results shown in Figures 6 and 7, which correspond to the observations we made for the VGG16 model, are accurate for the DenseNet and InceptionNet models, respectively. In both situations, we discovered that skin types that were closest to the training set had the greatest accuracy scores, demonstrating the biases in DNN models towards particular skin tones.

When assessing the InceptionNet model, we did find one anomaly in the FST 3-6 holdout set. Even though FST 5 was closer to the training set, we discovered that FST 6 had a higher accuracy score. Because there is only one anomaly in FST 6 and there aren't many images of it, we should be aware that we could ignore this one. If the model is retrained and reevaluated, this anomaly might also be fixed.

In conclusion, the accuracy patterns of the DenseNet and InceptionNet models were comparable to those of the VGG16 model, indicating that the biases of DNN models towards particular skin tones should be taken into consideration when training them on medical imaging datasets.

EXPLORED APPROACHES

We have explored three approaches to reduce the biased outcome of the DNN models.

1. Balancing the three partition labels:

We saw that there were notable imbalances in the data when we first visualized it in terms of the three division labels. We tried to amplify the data to balance the three labels in an effort to solve this problem. Why must the three division labels be balanced, one could wonder. Due to the fact that each of the three partition labels is a category of skin condition, if one label considerably outnumbers the others, it indicates that the conditions under that label are more prevalent in the dataset. In order to ensure that conditions with fewer photos are also included in the dataset, we have added two additional labels to the other two. It is crucial to understand that this augmentation process does not always take place uniformly or stratified, which is what would be ideal. Theoretically, this is correct, but there are also limitations and things to keep in mind.

Result: In order to evaluate the performance of our VGG16 model with that of the original dataset, we applied it to the expanded dataset without placing any restrictions on the training or testing data. Surprisingly, we discovered that incorporating the supplemented dataset considerably decreased the model's overall accuracy. This discovery led us to the conclusion that this data augmentation technique would not be very helpful.

2. Balancing actual skin conditions (labels) :

In terms of the number of images for each of the 114 skin condition labels, we observed a significant imbalance in the dataset. In particular, the first 20 labels each contained fewer than 50 images, whereas the remaining 94 labels contained 50 or more images. To remedy this, we increased the number of images for the first 20 labels, giving them a total of 50 images each. To create a new balanced dataset with a total of 5700 images, we also randomly picked 50 images from each of the remaining 94 labels. In particular, for the underrepresented skin conditions with fewer images, this balanced dataset can enhance the performance of deep learning models in accurately classifying skin conditions.

Result: The performance of the VGG16 model fell short of the original dataset after the data were balanced by enhancing the images. This could be for a number of reasons, including an insufficient augmentation process or the VGG16 model being better suited to the original dataset's inherent imbalance. Regardless, the outcome emphasizes the value of careful thought when balancing datasets and the need for testing to find the best method.

3. Changing contrast of lighter skin images:

We attempted to make the light-skin images appear darker by manually adjusting their contrast in order to address the issue of light-skin images being underrepresented in the dataset. With the help of the VGG16 architecture, we trained two models. The initial model was tested on FST5 and FST6 images after having been trained on FST1 and FST2 images. The second model was

tested on FST5 and FST6 images after being trained on FST1 and FST2 images with 50% less contrast.

Result: The tests revealed no discernible difference in accuracy between the two models with various contrast levels. This suggests that the issue of accuracy disparities in DNN models is not effectively resolved by artificially altering the contrast of the images.

CONCLUSIONS

Due to the models' sensitivity to skin tone, using deep neural networks to classify medical images may produce biased results. While it may seem logical to add more data to balance the number of images across different skin tones, our study has shown that this may not always be the best course of action. Even worse, altering skin tones artificially by adjusting contrast may cause new problems or skew representation. Alternative approaches are therefore required to address these biases and guarantee fair representation among various populations. To learn more about the characteristics and elements influencing bias in DNN models, future research might investigate more complex techniques for data augmentation as well as the application of explainable AI techniques. To ensure fair and accurate diagnosis and treatment for all patients, regardless of their skin color or other demographic factors, a multi-pronged approach involving data collection, preprocessing, and model development will ultimately be necessary.

REFERENCES

Groh, M., Sarin, K. Y., Cha, K. H., Sun, J., Han, S. S., Laga, H., & Joon Lee, J. (2020). Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. *Scientific reports*, 10(1), 1-11. doi: 10.1038/s41598-020-70686-2

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.

Food for Skin. (2021). Know Your Skin: What's Fitzpatrick Skin Type?. Food for Skin. <https://foodforskin.org/know-your-skin-whats-fitzpatrick-skin-type/>

Xu, Q., Song, Y., & Cheng, Y. (2019). An Intelligent Classification Model for Surface Defects on Cement Concrete Bridges. *Advances in Materials Science and Engineering*, 2019, 1-14.

draw.io. [Computer software]. Retrieved from <https://www.draw.io/>