Classifying Skin Tone on Skin Disease Datasets to Improve Fairness

G025 (s2447930, s2602230)

Abstract

There has been a drastic increase in the use of deep learning to classify skin diseases from medical images, achieving some impressive results. However, in many datasets darker skin tones are underrepresented, and this can lead to biased models that perform better on lighter skin tones. This perpetuates and amplifies unfairness towards marginalised groups, especially important in a medical setting where patient health is at risk. Most skin disease datasets do not have skin tone labels, making identification and mitigation of bias difficult. We propose a deep learning method to classify skin tone of skin disease images on the six-point Fitzpatrick scale. We improve upon the benchmark with an approach inspired by a model used in plant disease classification. This architecture uses the LAB colour space and processes brightness and colour in separate branches before an inception block considers data at different scales. We improve on the baseline by 4.5%, getting an accuracy of 48.1% and a give-one-accuracy of 86.9%. The approach can can classify skin tone on skin disease datasets without skin tone labels, helping identify bias in both datasets and models. It is more useful as a diagnostic tool than for detailed analysis, as the give-one-accuracy is much higher than the exact accuracy.

1. Introduction

1.1. Background

Skin diseases are a very broad group of conditions affecting skin, ranging from diseases like acne to melanoma, encompassing many different symptoms (Karimkhani et al., 2017). Skin diseases are an extremely common reason to see a medical professional, with 85 million Americans seen by a physician for skin diseases in 2013, and this number is expected to increase due to an aging population (Lim et al., 2017). This means that skin conditions have a big impact on both public health and resources of healthcare institutions (Lim et al., 2017).

The use of deep learning has been explored to alleviate some of the vast impact of skin conditions, especially in the field of medical diagnosis (Li et al., 2021), with the potential benefits of speeding up disease detection and increasing diagnostic accuracy. As the similarity of lesions caused by different skin conditions make diagnosis very hard (Li et al., 2021), an effective deep learning classification would relieve strain on the resources of healthcare institutions and improve patient outcomes (Li et al., 2021), (Lim et al., 2017). Furthermore, using deep learning to diagnose skin conditions has led to some promising results, with models even outperforming diagnosis by board-certified dermatologists (Fujisawa et al., 2019).

Some papers have raised concerns about lack of representation of darker skin tones in skin disease datasets and what this means for deep learning models trained on them (Guo et al., 2022). As the vast majority of data for skin disease datasets is collected from countries where the population is predominantly lighter skinned- like the Unites States, Australia and many European countries, datasets tend to have far more examples of skin diseases on lighter skin tones than those on darker skin tones (Guo et al., 2022). This has resulted in training of biased deep learning models, which perform better on the lighter skin tones that make up the majority of training data (Groh et al., 2021). If these biased models are used in practice they will contribute towards inequality by making marginalised groups with darker skin tones less likely to get the correct medical treatment.

This is very important because skin conditions do appear differently for different skin tones, significantly impacting diagnostic accuracy. For example, Lyme disease has a very atypical representation on darker skin tones compared to people with lighter skin tones (Ooi et al., 2021), highlighting a critical gap in medical data representation. This discrepancy is largely caused by the under-representation of patients with darker skin tone in datasets, meaning such diseases could be misdiagnosed and not treated on time. In the USA around 47% of the dermatologists experience difficulties when diagnosing skin diseases on darker skinned patients (Narla et al., 2023). Integrating deep learning methods into the diagnostic process could offer a solution, enhancing dermatologists' ability to accurately identify and treat skin diseases across a diverse range of skin tones.

1.2. Aims

These issues are exacerbated by most skin disease datasets not containing skin tone labels (Tschandl et al., 2018), (Codella et al., 2019), resulting in biased models being harder to detect, and a lack of awareness of the issues caused by imbalanced data. This paper aims to address this issue by training a model to classify skin tone. Specifically a two branch convolutional nerual network (CNN) is trained on the Fitzpatrick 17k dataset which has skin

tone annotations (Groh et al., 2021). These annotations are commonly used and on a six-point scale going from light skin to dark skin (Groh et al., 2021), the skin tones can be seen in Figure 1.

A deep learning model trained to reliably classify skin tone from skin disease images can then be used to label the attribute on datasets that don't have skin tone labels, aiding detection of bias in models trained on these unlabelled datasets. The performance of this model could further be improved by fine tuning on other datasets, which requires much less time and resources than annotating a full dataset with skin tone labels.

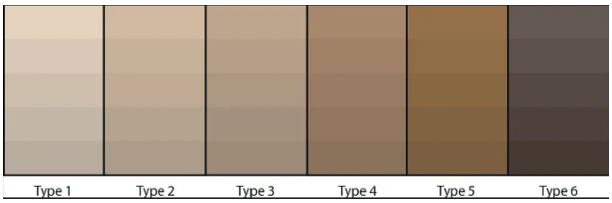

Figure 1. The six point Fitzpatrick scale for skin tone, ranging from 1 (lighter skin) to 6 (darker skin), image from (Tadesse et al., 2023)

1.3. Previous Approaches

Despite the vast increase in the use of deep learning models to classify skin disease images, there has been limited literature published on the related task of classifying skin tones on skin disease images (Bevan & Atapour-Abarghouei, 2022). The most common approach to classifying skin tone for use in various tasks, including bias detection, is by using the individual typology angle (ITA) (Groh et al., 2021). ITA is a non-machine learning method that applies a simple operation to pixel values in the LAB colour space to get a score based on how light skin tone is (Del Bino & Bernerd, 2013). Due to its simplicity and wide use this method is selected as a baseline to our experiments (Groh et al., 2021), (Tadesse et al., 2023). As far as we are aware there is only one paper that used a deep learning approach to classify skin tone, Tadesse et al. (Tadesse et al., 2023) carry out skin tone classification of skin disease images to quantify bias in dermatology textbooks. They also train on the Fitzpatrick 17k dataset, but instead do binary classification (classes 1-4 vs classes 5-6). The model they use is a fine-tuned ResNet-18 that was pre-trained on ImageNet, we use this model as the benchmark model for our experiments (Tadesse et al., 2023).

Due to the lack of deep learning methods for skin tone classification on skin disease images, we create an architecture based on a state-of-the-art-model for classifying plant disease (Schwarz Schuler et al., 2022). As plant disease images look very similar to skin disease images, as seen in Figure 2, and the colour of images is very important, there is a lot in common between these tasks.

Figure 2. Visual comparison of plant disease images from the plant village dataset (G. & J., 2019) and skin disease images from the Fitzpatrick 17k dataset (Groh et al., 2021). It can be seen that the shapes and colours of diseased areas in both datasets are similar, furthermore, both datasets have images that include background, diseased areas and non-diseased areas.

1.4. Contribution

The model we implement is a CNN that uses images converted to the LAB colour space. The LAB colour space uses three channels, L for lightness, and AB for colour, where AB is for the four unique colours of human vision, which are yellow, red, green and blue (Luo, 2014). The model has two branches, one which takes as input the L channel, and the other which takes as input the AB channels. By operating on each channel separately before combining for further processing the model can better learn how brightness and colour combine to predict skin tone (Schwarz Schuler et al., 2022). Furthermore, the model uses an inception block which processes the image at different scales, considering finer details and broader patterns helps in isolating skin and predicting tone.

Our model has an accuracy of 48.1% and a give-one-accuracy of 86.9% (where a prediction is marked as correct if it is one off the actual label). Our model is able to improve on the benchmark by 4.5% despite having less than 50% of the parameters the benchmark has, evidencing the benefit of the architecture. This model is therefore a better model

Dataset skin tone distribution						
Skin tone type	1	2	3	4	5	6
Images	2941	4796	3297	2775	1527	628

Table 1. The distribution of different skin tone types in the dataset. Lighter skin tones (1-3) are far more common than darker skin tones (4-6)

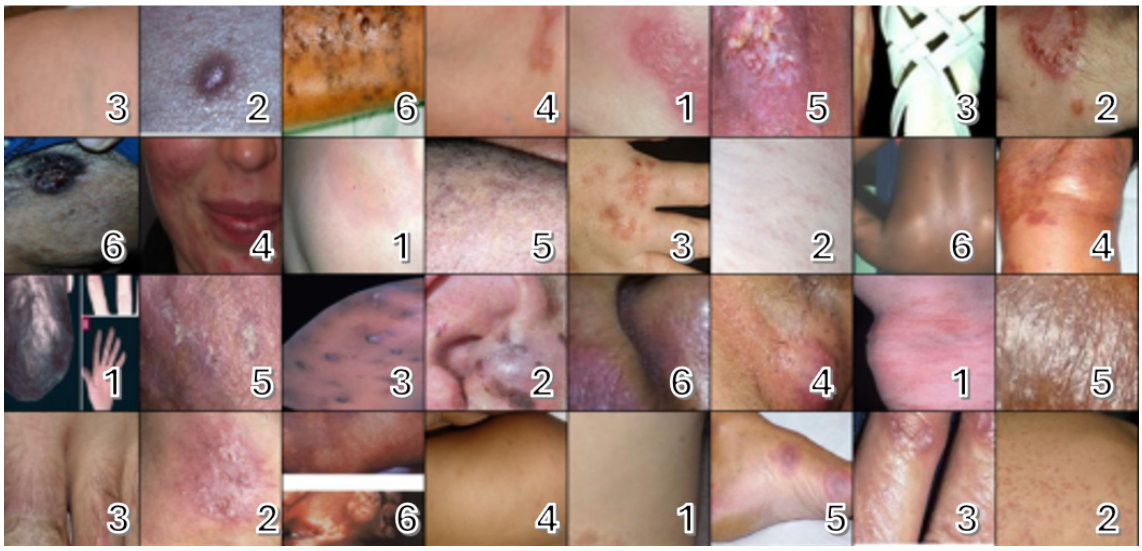

Figure 3. A batch of skin disease images from the Fitzpatrick 17k dataset (Groh et al., 2021). The numbers on each image indicate their skintone label

for predicting skin tone and diagnosing bias. However, the model performance is still modest, especially when trying to predict the exact category, therefore the model is more useful as a diagnostic tool than a tool for doing detailed analysis.

2. Data set and task

2.1. Data Set

We use the Fitzpatrick 17k dataset of skin disease images (Groh et al., 2021). The dataset contains around 17,000 images labelled with 114 skin conditions, as well as the skin tone of the patient on a scale of 1-6, with 1 being light skin and 6 being dark skin. Skin tone labels were annotated by a team of human annotators, with disease labels from open source dermatology atlases, of which 3% have been checked by a board certified dermatologist (Groh et al., 2021). Before utilizing the data any invalid urls or unlabelled images (where annotators could not agree) were removed. This left 15964 images, the remaining data was split into training and evaluation, where the validation data is a random 10% set of the whole data. A sample of images from the dataset is shown in Figure 3.

2.2. Unbalanced Data and Preprocessing

The dataset is unbalanced, with lighter skin tones being far more common than darker skin tones, the total number of images per each skin tone type is displayed in Table 1. We can see from the table that the lightest skin tone 1 has over four times as many images as the darkest skin tone 6, whilst the second lightest skin tone 2 has over seven times more images than skin tone 6.

To ensure models we train aren't biased towards lighter skin tones which have more representation, we over-sample any underrepresented skin tones in the dataset such that there are equal numbers of images for every skin tone. Furthermore, it is ensured that every batch also has a roughly equal number of images for every skin tone so training is consistent.

Alongside oversampling, we also apply further data preprocessing and augmentation techniques to increase the amount and diversity of images that our model sees. This is especially important to improve performance for our task as the dataset is not very big and some classes have a very limited amount of examples (Shorten & Khoshgoftaar, 2019). Our pre-processing includes resizing all images to 256x256, taking random crops of images of size 224x224, random horizontal flipping, random vertical flipping and finally normalisation.

2.3. Task

The task we train models our models for is multi-class classification of skin tone. Models output a probability of an image belonging to each of the six skin tone types, where 1 corresponds to lighter skin tones and 6 to darker ones. The maximum of these probabilities is taken as the

prediction.

This is a difficult task, images are diverse and cluttered with noise from the background and diseased areas. Furthermore, some classes appear very similar and it is difficult to perfectly classify many images into a skin tone label. For this reason, we evaluate all methods using accuracy and give-one-accuracy, where give-one-accuracy marks a prediction as correct if it is one higher or lower than the actual label. This is a common approach in the literature (Groh et al., 2021), (Tadesse et al., 2023). Getting the prediction close to the actual label is still beneficial in our task, as close predictions can still be used to identify bias, for example between lighter categories (1-3) and darker categories (4-6). We also analyse the accuracy per skin type for all our models to ensure they are not biased towards any particular types of skin tone. Models more accurate on certain skin tones could lead to incorrect diagnosis of biased data sets and models, which has the potential to negatively impact marginalised groups.

3. Methodology

3.1. Baseline

The baseline we use for skin tone classification is the individual typology angle (ITA). This is a fast and simple non-machine learning method to classify skin tone that is widely used in the literature (Groh et al., 2021), (Tadesse et al., 2023). The ITA value represents how dark or light a skin tone colour is, with higher values corresponding to lighter skin tones (Del Bino & Bernerd, 2013). Furthermore, its efficacy has been tested by comparing to pigmentation analysis using lab-based staining methods (Del Bino & Bernerd, 2013).

To calculate ITA value, images are first converted into the LAB colour space which is based on how humans see colour (Luo, 2014). The L channel, representing luminance, and the A channel, representing the yellow/blue component of colour, are used to calculate the ITA value. The formula for calculation is shown in equation 1.

$$ITA = \frac{180(\frac{arctan(L-50)}{B})}{\pi} \quad (1)$$

Where L is the average value in the luminance channel, and B is the average value in the yellow/blue channel. Values that are zero or over one standard deviation from the mean are discarded to improve reliability. The skin tone of an image is then classified by comparing the ITA values to thresholds. We use the common thresholds calculated specifically for dermatology in (Kinyanjui et al., 2020):

- ITA > 55 –> Skin Tone 1 (Very Light)
- 55 ≥ ITA > 41 –> Skin Tone 2 (Light)
- 41 ≥ ITA > 28 –> Skin Tone 3 (Intermediate)
- 28 ≥ ITA > 19 –> Skin Tone 4 (Tan)
- 19 ≥ ITA > 10 –> Skin Tone 5 (Brown)
- 10 ≥ ITA –> Skin Tone 6 (Dark)

3.2. Benchmark

Our benchmark is a pre-trained ResNet18 model as in (Tadesse et al., 2023). In this paper Tadesse et. al carry out skin tone classification for the purpose of quantifying bias in dermatology textbooks. They use a binary classification model (skin tones 1-4 vs skin tones 5-6) that is pre-trained on ImageNet and then fine-tuned on the Fitzpatrick 17k dataset (Tadesse et al., 2023).

As the task is highly related it is a suitable benchmark for our model. Our task is multi-class classification instead of binary classification so we alter the architecture from (Tadesse et al., 2023), changing the number of output logits to six and adding a softmax so the model outputs a probability for each class. As the task is multi-class classification, cross entropy loss is used. We also tested the effect of simple HSV and YCbCr threshold masking as in (Tadesse et al., 2023), but found it did not improve performance so did not include the results in our analysis.

3.3. Two-Branch CNN Model architecture

We present a deep learning model inspired from a state-of-the-art algorithm for plant disease detection (Schwarz Schuler et al., 2022) as it tackles a similar problem . Initially, the method takes a batch of images which have been transformed into the LAB color space. These images are subsequently divided into two separate images (img_L and img_AB): one containing only the L channel (img_L), representing the lightness from the original image and one comprising of the chromatic information through the AB channels. This separation of the images is utilized to inspect specific features hidden in the lightness and chromatic details of the original composition, allowing a more detailed analysis of the visual characteristics that might not be detected in the original images.

The img_L data is then passed to a convolution block (Conv 1) containing three 2D convolutional layers with 5, 5 and 11 output channels respectively and a MaxPool layer, each with 3x3 convolutional filters and no padding. There is a stride of 2 in the first convolution layer and the max pooling layer, and a stride of 1 for the rest. In a parallel manner, the img_AB images are passed to another convolutional block Conv 2 which mirrors the structure of the first one. The notable difference lies in the number of channels output by the three convolutions in the mirrored layer, which are set to 27, 27 and 53 respectively.

Consequently, the outputs of both convolution blocks are combined and inputted into another convolutional block (Conv 3). Conv 3 includes two 2D convolutional layers C1 and C2 both with stride 1 followed by a max pooling layer with stride 2. C1 has a 1x1 convolutional filter, 80 output channels and is zero padded while C2 contains 3x3 convolutional filters, 192 output channels and is not padded.

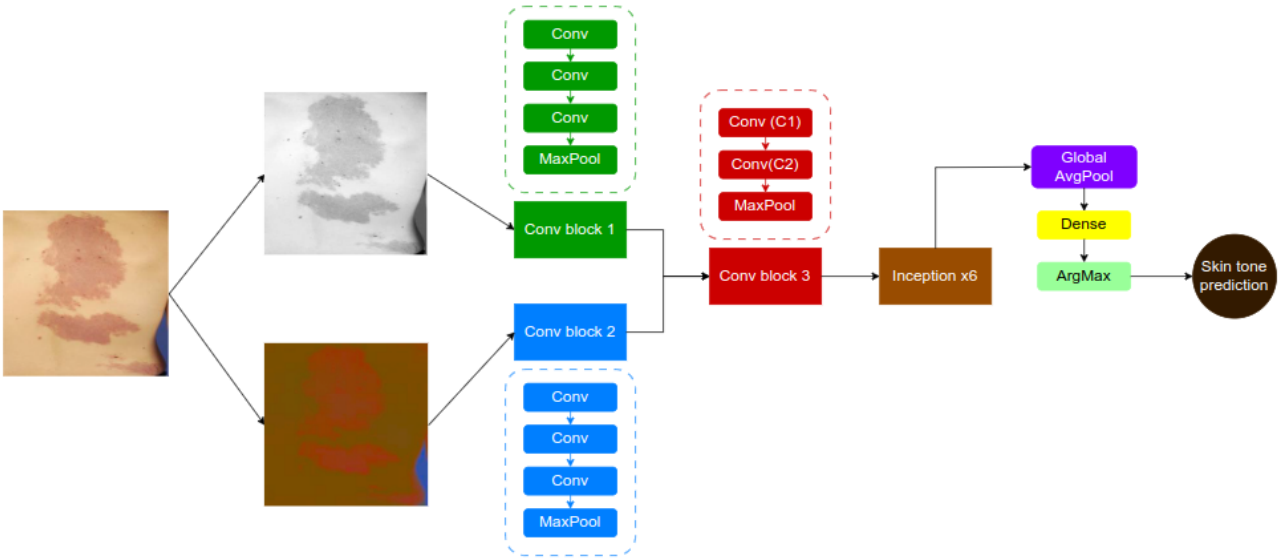

Figure 4. A diagram of the architecture of the skin tone detection algorithm. It visualize the steps needed for predicting a skin tone type from an image.

The process continues by passing the transformed data into an inception x6 block (modified V3 inception). The inception block captures and combines information about the data at different scales by using parallel convolution blocks with different kernel sizes. A more detailed explanation of the inception block structure can be found in Figure 5. The inception block generates data with 768 output channels which is passed into a global averaging pooling layer followed by a dense layer that narrows down the outputs to 6 (the number of skin tones). In the end, the most probable skin tone type is chosen. A diagram of the complete architecture can be seen in Figure 4.

All convolutional layers in the network are followed by a batch normalization and a ReLU activation function. The considered batched size is 32 and the loss that is used is cross-entropy.

3.4. Two-Branch CNN Hyper-parameters

The channels for the network were tuned in a similar way to the algorithm presented in the (Schwarz Schuler et al., 2022) paper. In a similar manner to the paper, parameters α and β control the number of channels in each branch, therefore specifying the relative importance of the L and the AB channels, where $\alpha+\beta = 64$. After experimenting we found that $\alpha = 11, \beta = 53$ achieves the optimal inclusion of the brightness and chromatic information needed for improving the skin tone detection performance.

4. Experiments

4.1. Baseline

The ITA value method produces an accuracy of 23.2%, and a give-one-accuracy of 53.2% on the Fitzpatrick 17k dataset. This accuracy is better than a random guess that would give an accuracy of 16.7%, but not high enough to reliably predict skin tone, even to within one skin tone type. Empirically selecting thresholds to maximise performance on Fitzpatrck 17k can improve the accuracy to 29.9% and the give-one-accuracy to 70.0%, which is considerably better but still not very reliable. ITA should only be used as a fast and simple baseline to classify skin tone of skin disease images

4.2. Benchmark

For our benchmark experiment we fine-tune a pre-trained ResNet18 model on the Fitzpatrick 17k dataset. Training is carried out until the validation performance hasn't improved for five epochs, at which point the best performing model is then saved.

Hyper-parameters were picked based off commonly used values in the literature. The hyper-parameters selected to train the benchmark were: batch size - 32, optimiser - stochastic gradient descent (SGD), SGD momentum - 0.9, SGD learning rate 0.001. Furthermore, exponential learning rate scheduling is chosen as this can aid model convergence (Smith, 2018), with parameters: step size - 10, gamma - 0.1. The model has around 11 trainable million parameters.

The benchmark ResNet18 model achieves an accuracy of 43.6% and a give-one-accuracy of 84.7% on the Fitzpatrick 17k dataset. This is considerably better than the baseline, improving on the accuracy by 20.4% and the give-one-accuracy by 21.5%. This proves the effectiveness of using deep learning methods for this task, and this model has good enough performance to be used as an initial diagnostic tool for finding bias, although not for more detailed analysis.

Analysis of the performance on individual skin types can be seen in Table 2. The model achieves the highest accuracy on skin tone type 1, with the lowest on type 3. The average accuracy on lighter skin tones (1-3) is 44.8%, and the av-

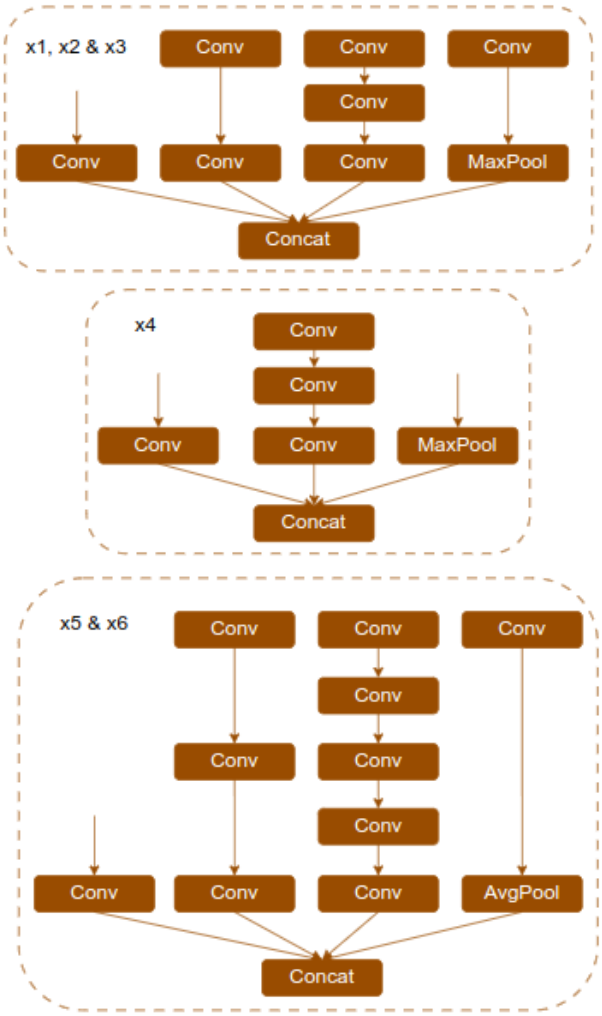

Figure 5. A representation of the architecture of the six layers in the Inception x6 block. The first three inception layers (x1, x2 and x3) share a similar architecture (displayed in the top figure). They simultaneously process the data with 4 different convolution paths, processing the data with 1x1, 5x5 and 3x3 convolution kernels as well as max pooling on the last path before combining them and moving to the next layer. The main difference between the three layers is the input channels which are 192, 256 and 288 respectively. The fourth inception layer (x4) has a slightly simplified structure which includes only 3 paths, two of which include 2D convolution layers each with 3x3 filters and one with only a max pooling layer. The last two inception layers (x5 and x6) have a similar structure (displayed in the bottom figure). All of the convolution paths in these inception layers contain a different number of 2D convolution layers with 7x7 filters. However, unlike the first four inception layers, the last path includes an average pulling layer. The only difference between x5 and x6 is the number of output layers that is 128 and 160 respectively.

Benchmark Accuracy on Different Skin Types						
Skin Tone Type	1	2	3	4	5	6
Accuracy (%)	65.6	41.5	27.2	43.1	39.0	46.2

Table 2. Accuracy of benchmark ResNet18 model on different skin tone types

Two-Branch CNN Accuracy on Different Skin Types						
Skin Tone Type	1	2	3	4	5	6
Accuracy (%)	70.9	42.6	34.9	38.3	45.9	50.0

Table 3. Accuracy of Two-Branch CNN model on different skin tone types

erage accuracy on darker skin tones (4-6) is 42.8%. These accuracies are similar, indicating the use of oversampling and pre-processing has mitigated biased predictions against darker skin types like those in (Groh et al., 2021).

4.3. Two-Branch CNN

The model we implemented to address our task is a Two-Branch CNN model based off the plant disease model from (Schwarz Schuler et al., 2022). More detail on how this model works can be found in section 3 (methodology) of this paper.

The parameters α and β described in section 3 that control the relative contribution of each branch of the Two-Branch CNN model were tuned to maximise performance. We found the best values to be 11 for α and 53 for β. All other hyper-parameters were set to be the same as those described in the Benchmark model section (4.2). The model has around 5 million trainable parameters.

The model achieved an accuracy of 48.1% and a give-one-accuracy of 86.9% which is the state-of-the-art for skin tone classification on skin disease images. The Two-Branch CNN is the best performing model, with an improvement of 4.5% accuracy and 2.2% give-one-accuracy over the benchmark, despite the model having less than 50% of the parameters of the benchmark. This indicates the effectiveness of the two branch architecture that separately operates on brightness, as well as the inception layers which consider data at different scales. The increase in accuracy is much greater than the increase in give-one-accuracy, suggesting that the model is better at distinguishing between close skin tones, but that there are still some samples that both models get completely wrong. These results make the Two-Branch CNN a promising tool to diagnose bias, especially if a detailed analysis isn't required, due to the high give-one-accuracy. However, the models performance isn't currently good enough to carry out a very detailed analysis of bias, as the accuracy is still below 50%.

Table 3 shows an analysis of the Two-Branch CNN model on different skin types. Just like the benchmark, the Two-Branch CNN performs best on skin type 1 and worst on skin type 3. This is likely because the very lightest skin tones are easiest to identify, whereas the skin tones in the middle of the scale are the most ambiguous. The Two-Branch CNN

Comparison of Different Methods				
Method	ITA (Kinyanjui)	ITA (Emperical)	ResNet18	Two-Branch CNN
Accuracy (%)	23.2	29.9	43.6	**48.1**
Give-one-accuracy (%)	53.2	70.0	84.7	**86.9**
Number of Parameters	N/A	N/A	**11M**	5M

Table 4. Accuracy, give-one-accuracy and number of parameters for different methods for skin tone classification on the Fitzpatrick 17k dataset. The number of parameters is rounded to the nearest million (M)

model has an average accuracy of 49.5% on the lighter skin types (1-3) and 44.7% on the darker skin types (4-6). The accuracy on lighter skin types is slightly higher, indicating that while over-sampling and pre-processing mitigate the issue, there is still a small amount of bias in the model towards lighter skin types. A summarised comparison of all methods can be seen in Table 4.

4.4. Ablation Study

We test the importance of different components of our method by removing them from the architecture and testing the effect on performance. The three main components that differentiate the Two-Branch CNN from the Benchmark ResNet18 model are:

1. Conversion of images to the LAB colour space
2. Use of separate branches for luminance and colour
3. Inception block

Therefore we test the effect of removing each of these in turn. When removing the inception block it is replaced with multiple convolutions to ensure the number of trainable parameters is similar to the original model. All models in this section were trained in the same way as the model in section 4.3, with identical hyper-parameters, loss function and training procedure.

Using the RGB colour space instead of the LAB colour space gives an accuracy of 46.7% and a give-one-accuracy of 86.4%, a 1.4% accuracy reduction and 0.5% give-one-accuracy reduction from the original model. This difference is modest, suggesting that while conversion to the LAB colour space makes it easier for the model to classify the exact skin tone type, the model can still learn to classify quite well from the RGB space.

Training a model with only one branch instead of two results in an accuracy of 45.2% and a give-one-accuracy of 85.7%. A reduction of 2.9% accuracy and 1.2% give-one-accuracy from the original model. This reduction is greater, showing the two branches are more important to the functioning of the model; however, the model is still able to beat the benchmark.

A model without the inception block achieves an accuracy of 43.0% and give-one-accuracy of 81.5%. A reduction of 5.1% accuracy and 5.4% give-one-accuracy. These results suggest the inception block is the most important part of the architecture for skin tone classification, with results lower than the benchmark when removed. The reduction is especially large in give-one-accuracy compared to the other ablations, possibly because consideration of the data at different scales helps the model to learn where healthy skin is, which reduces the number of completely wrong predictions. A summary of the performance of the ablations can be observed in Table 5.

Accuracy of Different Ablation Studies				
Ablation (%)	Original	RGB images	One Branch	No Inception
Acc (%)	48.1	46.7	45.2	43.0
Give One Acc (%)	86.9	86.4	85.7	81.5

Table 5. Accuracy and give-one-accuracy of different ablation studies

5. Related work

Our approach is based on a model architecture used in plant disease classification (Schwarz Schuler et al., 2022), which uses two branches to separately process brightness and colour, as well as inception layers to consider the data at different scales. We chose this architecture as plant disease classification is a similar task to skin tone classification, images contain background, healthy and diseased areas, and the colour is very important (G. & J., 2019). Given the success of our model, adapting other model architectures from the domain of plant disease (Lili Li, 2021) could yield good results for skin colour classification. Other related tasks that use methods with the potential to improve performance in skin tone classification include colour classification (Santosh Kumar et al., 2022) and skin disease classification (Li et al., 2021).

As far as we are aware, the only other paper performing skin tone classification on the Fitzpatrick 17k dataset is (Tadesse et al., 2023), that used a pre-trained ResNet18 model to carry out binary classification of skin tone. Therefore we use this method as the benchmark for our implementation. Despite the lack of papers on skin tone classification, there is a lot of literature on classifying skin disease on the Fitzpatrick 17k dataset (Li et al., 2021). Our task is motivated by the fact that many of the models trained to classify skin disease, along with the datasets they are trained on are biased towards lighter skin types (Groh et al., 2021). There is an increasing number of papers focusing on be-biasing classification of skin disease predictions (Reimers et al., 2021), (Nagpal et al., 2020), but not many papers on diagnosing the problem for unlabelled datasets or pre-existing models,

which our method aims to do.

Ensuring unbiased predictions is also important for our method, especially considering the Fitzpatrick 17k dataset is unbalanced. We tackle the problem by oversampling and data pre-processing, as mentioned in section 4.3 this mitigates but does not entirely remove bias. Applying different approaches to deal with unbalanced data has the potential to improve both prediction accuracy and fairness. This could be done by altering the model like mentioned in the previous paragraph (Reimers et al., 2021), (Nagpal et al., 2020), or via different data augmentation techniques (Sharma et al., 2020). There are also more advanced approaches like using generative adversarial networks (GANs) to generate more data. GANs have actually been used to vary the skin tone of data in different applications, so could be used to generatre more samples of minority classes for skin tone classification. (Roy et al., 2020).

6. Conclusions

Skin diseases are widespread category of health concerns that place a significant strain on public health and resources (Lim et al., 2017) and this strain is made worse by skin diseases being very hard to diagnose. Deep learning has shown promising results in automatically diagnosing skin diseases from images (Li et al., 2021), but many datasets and models used for this task are biased against darker skin tones (Groh et al., 2021). These issues are hard to detect and mititgate because most skin disease datasets do not contain skin tone labels. In this paper we train a deep learning model to classify skin tone on the Fitzpatrick 17k dataset to test whether a deep learning model can effectively label skin tone on skin disease images.

Our model is a Two-Branch CNN that separately focuses on brightness and colour, with inception layers to consider data at different scales. Our model beats the benchmark (current state-of-the-art) by 4.5% on accuracy and 2.2% on give-one-accuracy despite having far less parameters, indicating the effectiveness of the architecture. An ablation study is carried out to test the relative importance of unique architecture components: converting images to the HSV colour space, using two branches and using an inception block. Although removing all these components degrades performance, the biggest difference is seen upon removal of the inception block, suggesting that processing data at different scales is very useful in classifying skin tone.

Although the accuracy of 48.1% represents an improvement from the baseline and the benchmark, it is still quite low, implying that our model is not suitable for detailed analysis of skin tones on unlabelled datasets. However the give-one-accuracy of 86.9% is much higher, indicating the model could still be useful for an initial estimate of the bias in either a dataset or machine learning model's predictions.

Future work could explore ways to improve classification of skin tone to make a more reliable tool. This could include different model architectures, potentially inspired by those used in related domains such as disease classification (Li et al., 2021) and colour classification (Santosh Kumar et al., 2022). Different pre-processing techniques could also be tested, for example altering the data augmentation (Sharma et al., 2020), adding segmentation of healthy skin through traditional or deep learning methods (Nirupama & Virupakshappa, 2023) or increasing the amount of under-represented data types by synthetic data generation using techniques like GANs (Roy et al., 2020).

References

Bevan, Peter J. and Atapour-Abarghouei, Amir. Detecting Melanoma Fairly: Skin Tone Detection and Debiasing for Skin Lesion Classification. In Kamnitsas, Konstantinos, Koch, Lisa, Islam, Mobarakol, Xu, Ziyue, Cardoso, Jorge, Dou, Qi, Rieke, Nicola, and Tsaftaris, Sotirios (eds.), *Domain Adaptation and Representation Transfer*, Lecture Notes in Computer Science, pp. 1–11, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16852-9. doi: 10.1007/978-3-031-16852-9_1.

Codella, Noel, Rotemberg, Veronica, Tschandl, Philipp, Celebi, M. Emre, Dusza, Stephen, Gutman, David, Helba, Brian, Kalloo, Aadi, Liopyris, Konstantinos, Marchetti, Michael, Kittler, Harald, and Halpern, Allan. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC), March 2019. URL http://arxiv.org/abs/1902.03368. arXiv:1902.03368 [cs] version: 2.

Del Bino, S. and Bernerd, F. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology*, 169(s3):33–40, October 2013. ISSN 0007-0963. doi: 10.1111/bjd.12529. URL https://doi.org/10.1111/bjd.12529.

Fujisawa, Y., Otomo, Y., Ogata, Y., Nakamura, Y., Fujita, R., Ishitsuka, Y., Watanabe, R., Okiyama, N., Ohara, K., and Fujimoto, M. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *British Journal of Dermatology*, 180 (2):373–381, February 2019. ISSN 0007-0963. doi: 10.1111/bjd.16924. URL https://doi.org/10.1111/bjd.16924.

G., Geetharamani and J., Arun Pandian. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Computers & Electrical Engineering*, 76:323–338, June 2019. ISSN 0045-7906. doi: 10.1016/j.compeleceng.2019.04.011. URL https://www.sciencedirect.com/science/article/pii/S0045790619300023.

Groh, Matthew, Harris, Caleb, Soenksen, Luis, Lau, Felix, Han, Rachel, Kim, Aerin, Koochek, Arash, and Badri, Omar. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology With the Fitzpatrick 17k Dataset. pp. 1820–1828, 2021. URL https://openaccess.thecvf.com/content/

CVPR2021W/ISIC/html/Groh_Evaluating_Deep_Neural_Networks_Trained_on_Clinical_Images_in_Dermatology_CVPRW_2021_paper.html.

Guo, Lisa N., Lee, Michelle S., Kassamali, Bina, Mita, Carol, and Nambudiri, Vinod E. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—A scoping review. *Journal of the American Academy of Dermatology*, 87(1):157–159, July 2022. ISSN 0190-9622, 1097-6787. doi: 10.1016/j.jaad.2021.06.884. URL https://www.jaad.org/article/S0190-9622(21)02086-7/fulltext. Publisher: Elsevier.

Karimkhani, Chante, Dellavalle, Robert P., Coffeng, Luc E., Flohr, Carsten, Hay, Roderick J., Langan, Sinéad M., Nsoesie, Elaine O., Ferrari, Alize J., Erskine, Holly E., Silverberg, Jonathan I., Vos, Theo, and Naghavi, Mohsen. Global Skin Disease Morbidity and Mortality: An Update From the Global Burden of Disease Study 2013. *JAMA Dermatology*, 153(5):406–412, May 2017. ISSN 2168-6068. doi: 10.1001/jamadermatol.2016.5538. URL https://doi.org/10.1001/jamadermatol.2016.5538.

Kinyanjui, Newton M., Odonga, Timothy, Cintas, Celia, Codella, Noel C. F., Panda, Rameswar, Sattigeri, Prasanna, and Varshney, Kush R. Fairness of Classifiers Across Skin Tones in Dermatology. volume 12266, pp. 320–329, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59724-5 978-3-030-59725-2. doi: 10.1007/978-3-030-59725-2_31. URL https://link.springer.com/10.1007/978-3-030-59725-2_31. Book Title: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 Series Title: Lecture Notes in Computer Science.

Li, Hongfeng, Pan, Yini, Zhao, Jie, and Zhang, Li. Skin disease diagnosis with deep learning: A review. *Neurocomputing*, 464:364–393, November 2021. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.08.096. URL https://www.sciencedirect.com/science/article/pii/S0925231221012935.

Lili Li, Shujuan Zhang, Bin Wang. Plant Disease Detection and Classification by Deep Learning—A Review | IEEE Journals & Magazine | IEEE Xplore, 2021. URL https://ieeexplore.ieee.org/abstract/document/9399342.

Lim, Henry W., Collins, Scott A. B., Resneck, Jack S., Bolognia, Jean L., Hodge, Julie A., Rohrer, Thomas A., Van Beek, Marta J., Margolis, David J., Sober, Arthur J., Weinstock, Martin A., Nerenz, David R., Smith Begolka, Wendy, and Moyano, Jose V. The burden of skin disease in the United States. *Journal of the American Academy of Dermatology*, 76(5):958–972.e2, May 2017. ISSN 0190-9622. doi: 10.1016/j.jaad.2016.12.043. URL https://www.sciencedirect.com/science/article/pii/S0190962217300166.

Luo, Ming Ronnier. CIELAB. In Luo, Ronnier (ed.), *Encyclopedia of Color Science and Technology*, pp. 1–7. Springer, Berlin, Heidelberg, 2014. ISBN 978-3-642-27851-8. doi: 10.1007/978-3-642-27851-8_11-1. URL https://doi.org/10.1007/978-3-642-27851-8_11-1.

Nagpal, Shruti, Singh, Maneet, Singh, Richa, and Vatsa, Mayank. Diversity Blocks for De-biasing Classification Models. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–9, September 2020. doi: 10.1109/IJCB48548.2020.9304931. URL https://ieeexplore.ieee.org/abstract/document/9304931. ISSN: 2474-9699.

Narla, Shanthi, Heath, Candrice R, Alexis, Andrew, and Silverberg, Jonathan I. Racial disparities in dermatology. *Archives of dermatological research*, 315(5):1215–1223, 2023.

Nirupama and Virupakshappa. Enhancing Skin Disease Segmentation with Weighted Ensemble Region-Based Convolutional Network. *Engineering Proceedings*, 59(1):49, 2023. ISSN 2673-4591. doi: 10.3390/engproc2023059049. URL https://www.mdpi.com/2673-4591/59/1/49. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Ooi, Rucira, Lim, Sheryl Li Xin, Ooi, Setthasorn Zhi Yang, and Bennett, Alistair. Representing black, asian and minority ethnic skin in dermatology education amidst the covid-19 pandemic: An evaluation of an e-learning resource. *Cureus*, 13(12), 2021.

Reimers, Christian, Bodesheim, Paul, Runge, Jakob, and Denzler, Joachim. Towards Learning an Unbiased Classifier from Biased Data via Conditional Adversarial Debiasing, March 2021. URL http://arxiv.org/abs/2103.06179. arXiv:2103.06179 [cs].

Roy, Debapriya, Mukherjee, Diganta, and Chanda, Bhabatosh. An Unsupervised Approach towards Varying Human Skin Tone Using Generative Adversarial Networks, October 2020. URL http://arxiv.org/abs/2010.16092. arXiv:2010.16092 [cs].

Santosh Kumar, N C, Uma Maheswari, S, M, Vigneshwari., Pramila, P V, Khilar, Rashmita, and Kumar, Ashok. Colour based Object Classification using KNN Algorithm for Industrial Applications. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 1110–1115, Pudukkottai, India, December 2022. IEEE. ISBN 978-1-66546-084-2. doi: 10.1109/ICACRS55517.2022.10029315. URL https://ieeexplore.ieee.org/document/10029315/.

Schwarz Schuler, Joao Paulo, Romaní, Santiago, Abdelnasser, Mohamed, Rashwan, Hatem, and Puig, Domenec. Color-Aware Two-Branch DCNN for Efficient Plant Disease Classification. *Mendel*, 28:55–62, June 2022. doi: 10.13164/mendel.2022.1.055.

Sharma, Shubham, Zhang, Yunfeng, Ríos Aliaga, Jesús M., Bouneffouf, Djallel, Muthusamy, Vinod, and Varshney, Kush R. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of*

the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20, pp. 358–364, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375865. URL https://dl.acm.org/doi/10.1145/3375627.3375865.

Shorten, Connor and Khoshgoftaar, Taghi M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL https://doi.org/10.1186/s40537-019-0197-0.

Smith, Leslie N. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, April 2018. URL http://arxiv.org/abs/1803.09820. arXiv:1803.09820 [cs, stat].

Tadesse, Girmaw Abebe, Cintas, Celia, Varshney, Kush R., Staar, Peter, Agunwa, Chinyere, Speakman, Skyler, Jia, Justin, Bailey, Elizabeth E., Adelekun, Ademide, Lipoff, Jules B., Onyekaba, Ginikanwa, Lester, Jenna C., Rotemberg, Veronica, Zou, James, and Daneshjou, Roxana. Skin Tone Analysis for Representation in Educational Materials (STAR-ED) using machine learning. *npj Digital Medicine*, 6(1):1–10, August 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00881-0. URL https://www.nature.com/articles/s41746-023-00881-0. Publisher: Nature Publishing Group.

Tschandl, Philipp, Rosendahl, Cliff, and Kittler, Harald. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, August 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL https://www.nature.com/articles/sdata2018161. Number: 1 Publisher: Nature Publishing Group.