DEPARTMENT OF ENGINEERING MATHEMATICS

CANCER MACHINE LEARNING MODEL TONE BIAS

Sohini Biswas

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Engineering in the Faculty of Engineering.

Friday 29th August, 2025

Supervisor: Dr. Ayush Joshi

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MEng in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

, Friday 29th August, 2025

Contents

List of Figures

List of Tables

Abstract

Skin cancer is considered one of the most common cancers in the world. Rapid development in the field of Artificial Intelligence is helping us transform how we detect and diagnose diseases. Deep convolution neural networks (CNN) have shown an excellent potential for data and image classification. In the field of dermatology, convolutional neural networks (CNNs) have been used to identify skin cancer from images. While AI tools have revolutionised many fields, it has lagged in the field of medicine due to technical challenges in developing solutions that cater to diverse population.

Many of the public datasets used to train these models, such as the open-source skin cancer image dataset from The International Skin Imaging Collaboration, contain far more images of lighter skin tones than darker ones. This is the result of clinical trials conducted in countries with lighter skin tones. Due to this tone imbalance, machine learning models derived from these datasets can perform well at detecting skin cancer for lighter skin tones. Though there is less prevalence of skin cancer with darker tones, any tone bias in these models would introduce fairness concerns and reduce public trust in the artificial intelligence health field.

The goal of the project is in two folds. Firstly, the project investigates whether CNN-based skin cancer classifiers exhibit skin tone bias. Using a subset of ISIC images with Fitzpatrick skin type annotations. To understand the model tone bias, a subset of images has been used which has more light skin tone images than dark skin tone images, and more benign than malignant images. The model will be trained with this imbalanced dataset and compare against a sample balanced dataset. The aim is to find if the model is biased in classifying the images as benign or malignant with respect to skin tone. Once it has been assessed whether the model is biased or not we move to the second part of the project to assess if other CNN models also give the same result. This will help in drawing a clear conclusion. Finally, we will add more images to see the effect of manual annotation and find out if other subgroups like gender also play a role in creating bias in the AI models

Ethics statement: This project fits within the scope of the blanket ethics application, as reviewed by my supervisor Dr Ayush Joshi.

Supporting Technologies

- Code was developed using a recent version of Python 3, utilising data science libraries such as NumPy,Pandas and matplotlib.
- Tensorflow , Keras are to be used for model development, training and testing
- I used Amazon Web Services for remote storage and processing of data. Specifically, I used:
  - Simple Storage Service (S3) for data storage
  - Elastic Compute Cloud (EC2) for provision of virtual machines
- I used LaTeX to format my thesis, via the online service *Overleaf*.

Acknowledgements

I would like to give my first and foremost thanks to my thesis supervisor Dr Ayush Joshi in the School of Engineering and Mathematics at the University of Bristol. He was always there to guide me when I had questions about the research or implementation problems. He was always forthcoming with promising resources and advice that allowed my work to reach its full potential. I also thank him for arranging minor financial support through the university to cover the costs of additional cloud computing resources. These resources enabled for a greater range of experiments to take place, enriching the results this project was able to achieve.

Finally, I must acknowledge the support provided by friends and family that carried me through the difficult and laborious periods of the project. Without their continual assurances this work would have been near insurmountable. You have my sincerest thanks.

Chapter 1

Introduction

In this chapter, the motivations for the thesis are outlined, followed by a justification into why research within this domain is critical. An overview of what this project seeks to investigate is then given, addressing the likely challenges involved. Finally, the key aims of the project and their derived objectives are described.

1.1 Motivation

Health disparities are one of the most pressing health issues we face today, transcending the boundaries of any single profession or discipline despite many decades of research and novel interventions [1]. Within the field of dermatology, such disparities are particularly evident in the diagnosis of skin cancer which is which is one of the most common cancers worldwide. Clinical evidences show that early detection is critical and it increases the survival rates. However, access to timely and accurate diagnosis is not available to everyone. People with colour are often diagnosed at later stages of disease progression, resulting in poorer prognoses and higher mortality rates.

AI is a rapidly developing field of study with incredible potential to change how we deliver care and provide health services. In particular, convolutional neural networks (CNNs) have demonstrated great performance in classifying skin lesions when trained on large annotated datasets such as those provided by the International Skin Imaging Collaboration (ISIC) [3]. Therefore, the aim of the thesis is to find the skin tone bias. The motivation is to understand how the CNN model performs. Transfer learning strategies which leverage pre-trained architectures such as AlexNet, VGGNet, and ResNet, have shown promising outcomes and reduced dependence on large medical datasets.

Despite these advances, fairness of the AI model remains a concern. Publicly available datasets are heavily imbalanced, containing disproportionately more images from lighter skin tones than darker skin tones. Consequently, models trained on such datasets often demonstrate reduced accuracy for darker tones. Such disparities not only compromise diagnostic reliability but also erode public trust in the adoption of AI technologies in clinical practice.

Motivation for this study is therefore in two folds. First, it seeks to assess whether CNN-based classifiers exhibit skin tone bias in skin cancer diagnosis which is in line with the work done in James Pope , Md Hassanuzzaman ,Mingmar Sherpa, Omar Emara, Ayush Joshi 1,and Nirmala Adhikari - SKIN CANCER MACHINE LEARNING MODEL TONE BIAS [3]. Second it explores transfer learning architecture to asses if the outcome is better with addition to exploring other parameters which can create bias in diagnosis like gender. By extending fairness analysis to both skin tone and gender, models can be assessed more comprehensively for equitable diagnostic performance.

In doing so, this research contributes to the growing body of work on vertical AI models which are domain-specific systems. Tailored to medical context, it will emphasize not only on predictive accuracy but also fairness and equity in clinical decision-making.

1.2 Challanges

- One issue with using Multilayer Perceptron to process image data is that they are not translation invariant. This means that the network reacts differently if the main content of the image is shifted. Since MLPs respond differently to shifted images, it complicates the classification process and produce unreliable results.[7] Convolutional neural networks (CNNs) address several limitations of traditional machine learning methods such as multilayer perceptrons (MLPs), their application to image analysis introduces its own set of challenges.

- One of the most significant challenges in medical image analysis is dataset imbalance. Publicly available datasets like ISIC archive are dominated with benign lesions and by images of lighter skin tones. This introduces biases in CNN models toward majority classes, resulting in high overall accuracy while masking poor sensitivity for malignant lesions and underrepresenting of darker skin tones. Such skewed learning undermines the reliability of AI in clinical settings.
- Medical datasets are usually small and require costly expert annotation. The scarcity of labelled data increases the risk of overfitting and reduces model generalizability. Transfer learning alleviates this issue but introduces dependency on features learned from non-medical domains, which may not always translate well to dermatology.
- Training CNN models require significant amount of computational power, including high-performance GPUs, large memory capacity, and optimized software libraries. Architectures like VGG-16 involve millions of parameters, demanding substantial training time and storage. To address these technical limitations, cloud-based pipelines are increasingly necessary for training, evaluating, and deploying these models.

1.3 Objective

The aim of this thesis is to investigate whether CNN-based skin cancer classifiers exhibit bias when trained on publicly available datasets.

- To design and implement CNN architectures for skin cancer classification which is in accordance to work done in *"SKIN CANCER MACHINE LEARNING MODEL TONE BIAS - James Pope, Ayush Joshi "* .
- To evaluate model performance of transfer learning CNN architecture using both traditional metrics as well as fairness metrics.
- To find out if other subgroups like gender also play a role in creating bias.

Do CNN-based skin cancer classification models exhibit bias across different skin tones and genders, and how can such bias be quantified and mitigated?

Chapter 2

Convolutional Neural Networks for Skin Cancer Detection: Foundation and Approach

This chapter introduces the fundamental concepts of CNNs, their architectural components, and the role of transfer learning in medical image analysis, with a particular focus on applications to skin cancer detection. By establishing this technical foundation, the chapter highlights why CNNs represent the most appropriate and effective approach for addressing the dual challenges of early diagnosis and fairness in healthcare AI.

2.1 Skin Cancer and the Need for Early Diagnosis

Skin cancer is among the most prevalent cancers worldwide, with incidence rates continuing to rise across populations. Clinical studies emphasize that early diagnosis is the most critical factor influencing prognosis, as treatment is significantly more effective when lesions are detected in their initial stages [17]. Despite the availability of dermoscopic imaging and improved diagnostic techniques, disparities persist in the timeliness of detection across different demographic groups.

These disparities are especially seen in individuals with darker skin tones. Although the incidence of skin cancer is lower in this group compared to lighter-skinned populations, it is often diagnosed at a later and more advanced stage, leading to worse outcomes.[18]. Several factors contribute to this discrepancy.

- Early visual indicators of skin cancer may be concealed by darker skin's higher melanin content. Atypical moles that would be more obvious on lighter skin tend to blend in with the surrounding skin tone due to subtle changes in pigmentation, texture, or appearance. Both patients and clinicians find it more difficult to identify suspicious lesions at first because of this.

- Darker skin tones are more likely to develop skin cancers in less sun-exposed areas, like the palms, nail beds, mucosal surfaces, and soles of the feet. Lesions are more likely to go unnoticed until they are advanced because these areas are not frequently examined during self-examination or even routine dermatological examinations.

- From a clinical perspective, darker skin tones can mask or obscure early dermatological changes, making lesion recognition more challenging. In addition, individuals with darker skin are often underrepresented in clinical studies and reference datasets used to train and validate these diagnostic tools. This under representation not only limits the expert in recognizing subtle presentations but also biases algorithmic models, as most have been developed and validated predominantly on lighter-skin tone datasets.

- Because the overall incidence of skin cancer is lower in individuals with darker skin, both patients and healthcare providers may underestimate the risk. This perception often results in fewer routine skin checks, reduced vigilance, and delayed consultations.

The consequence is a systematic skew in diagnostic accuracy and sensitivity across populations. While technological innovations in imaging and machine learning hold promise, their equitable application

requires deliberate efforts to diversify training datasets, improve clinical representation, and develop methods robust to variations in skin tone. Addressing these inequities is essential to reduce late-stage diagnoses, ensure fair access to accurate diagnostic technologies, and ultimately improve outcomes across all demographic groups.

2.2 CNN and its Background

David Hubel and Torsten Wiesel were two neurophysiologists who experimented in 1959 and eventually published their findings in a work titled "Receptive-Fields of Single-Neurons in cat's straits cortex" [13]. Their work defined how the neurons in a cat's brain are organized in a tiered pattern or layered form. These are the layers that can learn to detect visual patterns with the help of local features, which are extracted first, and for a higher-level representation, the extracted features are then combined. This concept has become the core behind the deep learning principles [14]. CNN is known for its ability to discover and interpret patterns. This property of pattern detection is useful in image analysis and hence heavily used in the field of medicine. CNN is a combination of layers which are classified into different categories.

Input layer: The raw data is stored in the first layer called the input layer. Convolutional Neural Networks take advantage of the fact that the input consists of images and they constrain the architecture in a more sensible way. In particular, unlike a regular Neural Network, the layers of a CNN have neurons arranged in 3 dimensions: width, height, depth.[8]

Convolution layer: A **convolutional layer** can be thought of as the "eyes" of a CNN. The neurons in a convolutional layer look for specific features. At the most basic level, the input to a convolutional layer is a two-dimensional array which can be the input image to the network or the output from a previous layer in the network. A convolutional layer is responsible for calculating the output volume by performing a dot product between the image patch and all of the filters, followed by another important function known as activation. The mathematical function is then applied to every element of the convolution layer's output [14]. A key feature of convolutional layers is called **parameter sharing**, where the same weights are used to process different parts of the input image. This allows us to detect feature patterns that are **translation invariant** as the kernel moves across the image. This approach improves the model efficiency by significantly reducing the total number of trainable parameters compared to fully connected layers. [7]

Pooling Layer: A pooling layer is a fundamental component in Convolutional Neural Networks (CNNs) used to reduce the spatial dimensions (width and height) of feature maps while retaining the most important information. This operation helps in dimensionality reduction, improving computational efficiency, and achieving translation invariance.

Activation Functions: Activation functions are used to transform an input signal into an output signal. This output signal is then used as input by the subsequent layer in the stack. The most common activations used in the study are ReLU function and Softmax activation function.

Dropout: Droupout is a regularization technique that is applied to the fully connected layers or convolutional layers in CNNs to prevent overfitting by randomly setting a fraction of activations to zero during training; this encourages the network to develop more robust features without relying too much on weights of specific neurons.[7]

Transfer Learning

Transfer learning is a concept where we train a model on one problem and then we can fine-tune and apply it on another similar kind of problem. Transfer learning is beneficial in terms of reducing training time, also need for huge datasets is eliminated. With the help of Transfer Learning, a model can be taught and refined for one activity and then applied to a different one which is closely connected to it. [10]

By keeping constant the baseline learning topology, various CNN architectures were proposed to improve the respective system performance. Among these, AlexNet, VGG16 and VGG19 are the famous CNN architecture introduced for object recognition task. [11]

Pretrained models like VGG16 are already trained on ImageNet which comprises of many categories of images. These models are built from scratch and trained by using high GPU's over millions of images consisting of thousands of image categories. As the model is trained on huge dataset, it has learned a good representation of low-level features like spatial, edges, rotation, lighting, shapes and these features

can be shared across to enable the knowledge transfer and act as a feature extractor for new images in different computer vision problems. These new images might be of completely different categories from the source dataset, but the pretrained model should still be able to extract relevant features from these images based on the principles of transfer learning. [12]

2.3 Literature Review

Clinical Background of Skin Cancer

Skin Cancer is one of the most common forms of cancer worldwide. Clinical research emphasis that early detection significantly increases the survival rates. As explained in the study by Kolm, Hofbauer Braun [17] that treatment is more effective if the lesions are identified at an early stage. This is why accurate tools are required for medical practice.

However, skin cancer is not evenly distributed across population. While cases of skin cancer are higher in Caucasians (light skin tone) the outcomes can be sever for people with darker skin tone due to delayed diagnosis. This was identified by Gloster and Neal (2006) [18] in their work where they highlighted that individuals with colour had advanced disease stages at diagnosis, resulting in poorer prognoses compared to their lighter-skinned counterparts. These disparities are due to lack of representative clinical datasets and tools primarily trained on light skin tones.

CNNs in the field of Skin Cancer

Application of Artificial Intelligence has seen tremendous growth in the field of medicine. One of the most critical fields of medicine – oncology has the most potential to benefit from it. Convolutional Neural Networks are being used for dermatological image analysis. CNN models like AlexNet and VGGNet established the foundation for deep feature learning (Wiesel, 1968 [13]; Sharma et al., 2020 [15]). Transfer learning, particularly with architectures like VGG16, has been successfully applied in dermatology to classify malignant images, where limited annotated datasets pose a challenge (Shaha Pawar, 2018 [11]; Tammina, 2019 [12]; Faghihi et al., 2020 [9]).

Challenges of Data Imbalance

Despite significant success, the performance of CNN models is affected by imbalance datasets. Most dermatology datasets contain more benign than malignant lesions and disproportionately represent lighter skin tones. This may be a reason that most of the research is conducted in western countries where the proportion of lighter skin tones are more. This has been shown to create predictive disparities, where models achieve higher sensitivity and specificity for light skin tones compared to dark tones (Ekellem Köhler, 2023 [4]). Under representation of darker tones in training data further limits model generalization. Public datasets such as the **ISIC Archive** [3] have been crucial in enabling this progress. ISIC Archive contains wide range of images from all skin tones.

Fairness of Machine Learning Models in healthcare

Fairness in a machine learning model is extremely critical and has become an important area of research. Suresh Guttag [6] in their work highlighted that bias can arise at different stages of machine learning lifecycle, from creating the dataset to training the model. This is why traditional performance measures such as accuracy, precision, recall are not enough parameters to find the bias.

Gender Bias in Skin Cancer Models

While a lot of research has been done towards bias related to skin tone, gender bias in skin lesion diagnosis has also emerged as a concern. Recent work on Evaluating Gender Bias and Fairness in Skin Lesion Diagnoses using Convolutional Neural Networks - Aakash Kondaka [19] shows that model performance can differ significantly across male and female subgroups, raising fairness concerns in clinical practice. Gender-based disparities can arise due to difference in dataset composition or during clinical annotation. This line of research emphasizes the importance of evaluating fairness along multiple demographic

dimensions and not only skin tone.

Transfer Learning for Medical Imaging

Transfer learning has been widely adopted to improve performance in medical imaging tasks with limited datasets. By leveraging pre-trained networks on large-scale image datasets, CNNs can capture low-level features (edges, textures, shapes) that are transferable to medical domains (Salehi, 2020 [14]). Studies have shown that transfer learning enhances performance, accelerates convergence, and reduces overfitting in small datasets (Shaha Pawar, 2018 [11]; Tammina, 2019 [12]). In dermatology, architectures like VGG16 and VGG19 have achieved strong results for skin cancer classification (Faghihi et al., 2020 [9]; Gupta et al., 2022 [10]).

Chapter 3

Understanding The Dataset

This chapter aims at providing a detailed explanation of the dataset used for training and examining the convolutional neural network models in the context of skin cancer classification. A thorough exploration of the dataset is the foundation of any study. This chapter will provide a detailed analysis of sourcing of the data, preprocessing steps and meaningfully creating a subset which will be used training the models. Furthermore, this chapter will also cover exploratory data analysis which will give insights how skin cancer is actually spread in the real world.

3.1 Sourcing Data

In this study, the dataset used is from International Skin Imaging Collaboration (ISIC) Archive, a widely recognized open-access repository of dermoscopic images for skin cancer research. [3]. To ensure the appropriateness of the dataset for skin tone bias analysis, only images annotated with Fitzpatrick skin type labels and diagnostic labels (benign or malignant) were included. The archive includes over 81,000 dermoscopic images which have been diagnosed as malignant or benign. Out of this huge set, 3,623 images have Fitzpatrick Skin Type (FST). This annotation will help determining whether the images are light or darker skin tone. The Fitzpatrick Skin Type (FST) annotation has been categorised into FSTI, FSTII, FSTIII, FSTIV, FSTV and FST IV. The set of images come with metadata which consists of much more information other than just Fitzpatrick Skin Type (FST) and diagnostic labels like age, gender, image id.

3.2 Preprocessing Data

It is the fundamental step in any machine learning project. It serves as a foundation for model development. Effective pre-processing helps in developing a clean dataset which will be then used as an input for the machine learning models. The archive includes not only the dermoscopic images but also their meta data. Proper cleaning and understanding of the metadata can give a lot of insights about the topic. It will not only help us train the models but also understand how is this disease spread across the globe.

1. Data Cleaning: The metadata includes a lot of information and all of it may not be required for the project. Few of the columns are selected like isic_id corresponding to the image ids, fitzpatrick_skin_type corresponding to the skin tone, diagnosis_1 which refers to melanoma or benign and sex. But this is not enough to draw insights from the dataset. Each Fitzpatrick type is mapped to a proper label for better understanding. After the labeling is done, each fitzpatric skin type is divided between two skin tones of light or dark. This decision is somewhat arbitrary. It will be the basis of judging the bias in the model.

- FST I, FST II → *light skin tone*
- FST III, FST IV, FST V → *Dark skin tone*

  Here the dataset has been created which can studied to draw insights and used as input for the models.

2. Dataset Splitting: Using the cleaned dataset, a balanced dataset is created which is discussed in section 3.4. It is then split into training and validation set following a 70:30 ratio respectively.

isic_id	fitzpatrick_skin_type	diagnosis_1	sex	FST_label	FST_Skintone
ISIC_6816081	II	Benign	Male	FST II	Light
ISIC_4068636	II	Benign	Male	FST II	Light
ISIC_1054769	III	Benign	Male	FST III	Dark
ISIC_9340537	II	Benign	Female	FST II	Light

Table 3.1: Imbalanced Dataset Metadata

3.3 Exploring The Dataset

3.3.1 Data Distribution based on Skin Type

The skin types have been divided into 5 categories – FSTI, FSTII, FST III, FSTIV, FST V and FST VI. The bar chart shows the count of images per Fitzpatrick Skin Type (FST) in the dataset. There is a clear imbalance in the dataset with FSTII having highest number of images with 47.1% share. This is followed by FST III and FST II with 16% and 12.3% images respectively. In contrast, darker skin types FST IV (8.98%), FST V (8.1%), and FST VI (7.42%) — are noticeably underrepresented.

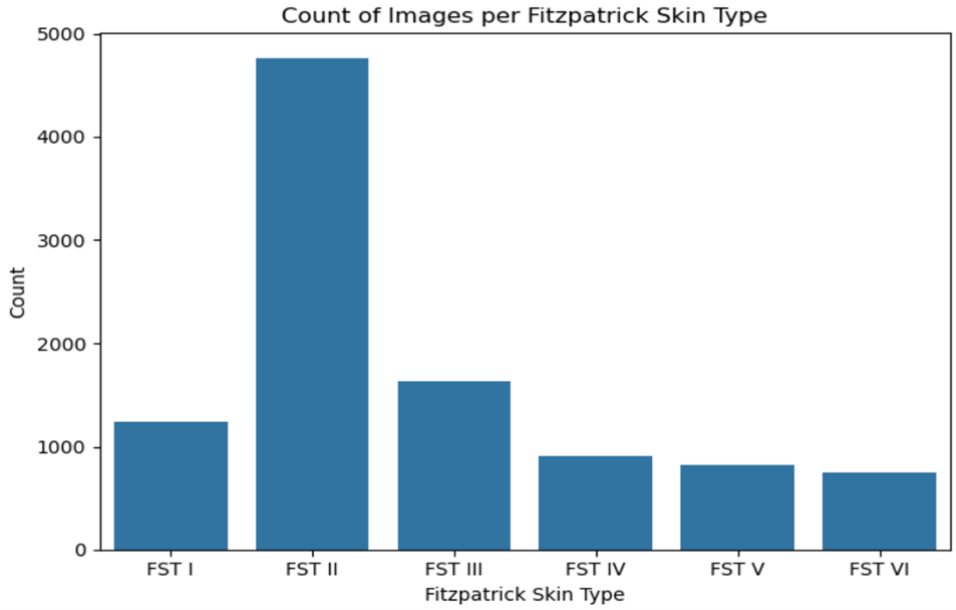

Figure 3.1: Image count based on FST scale

3.3.2 Data Distribution based on Diagnosis

The dataset shows a clear class imbalance between the diagnosis of benign and malignant images. The bar graph clearly shows that benign cases dominate with 8,055 images, while malignant cases account for only 2,066 images. The dataset exhibits two significant sources of imbalance: skin tone distribution and diagnostic class proportions.

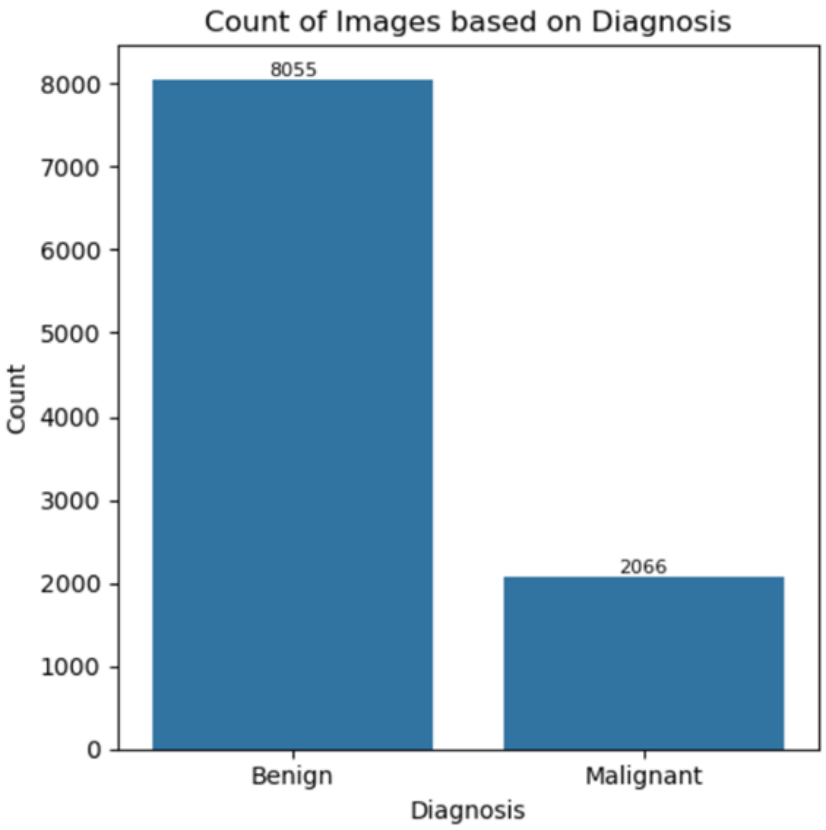

Figure 3.2: Image count based on diagnosis

3.3.3 Data Distribution based on Gender

To further understand the dataset, I explored how the diagnosis spread across both the genders. Few of the images don't have the sex annotation. 5,301 images are of female while 4,788 are male. Out of this 17.88% females have been diagnosed as malignant and 23.26% were males. (Figure 3.3)

The second stacked bar chart further decomposes this relationship by incorporating skin tone. It reveals that light-skinned individuals are overrepresented in both benign and malignant categories across genders, particularly among females. In contrast, darker-skinned samples are underrepresented—especially in the malignant category for both the genders. Least number of representation can be seen for dark and malignant cases across both the genders. The dataset has more number of images from female patients as compared to male patients but the patern of bias is similar in both the genders.

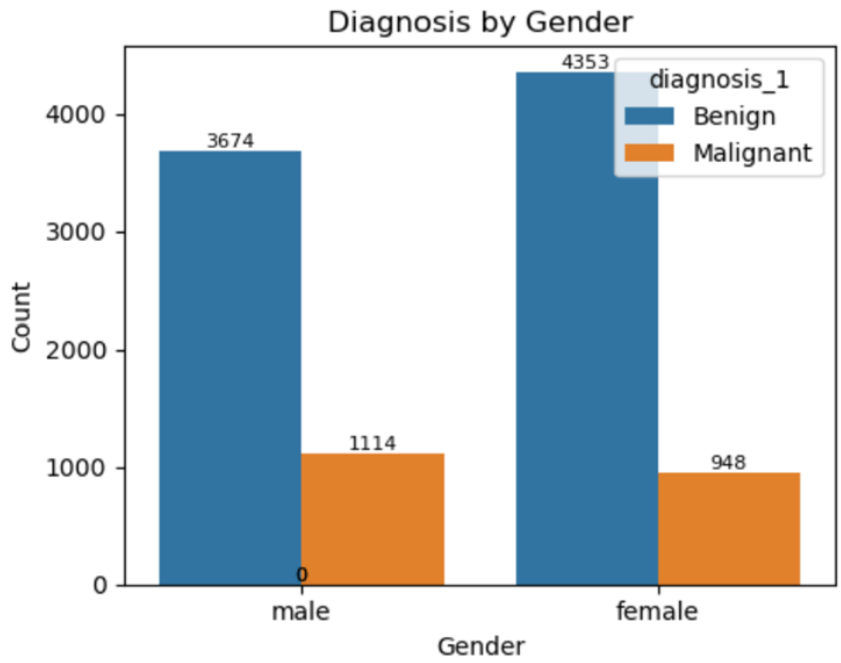

Figure 3.3: Diagnosis by Gender

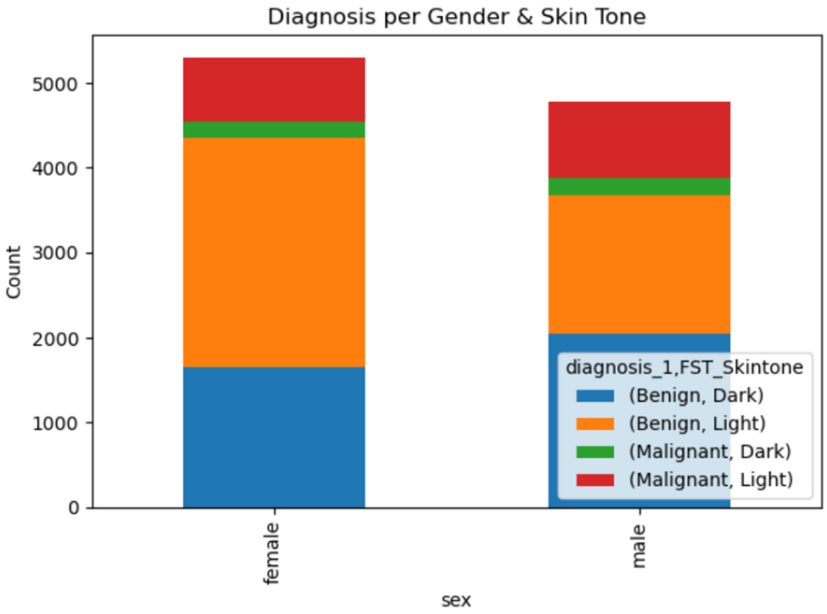

Figure 3.4: Diagnosis per Gender and Skin Tone

3.4 Lack of Image Diversity

The dataset obtained from the archive represents substantial imbalances. The archive has 500,000 images out of which 10121 images have been labelled as Benign and Malignant and have been categorised under different skin types. So, we can use these 10121 images for training our models. But even this has some drawbacks. The dataset exhibits two significant sources of imbalance: skin tone distribution and diagnostic class proportions. From the figure below, we can observe that there is a strong skew in the representation of Fitzpatrick Skin Types (FST), with lighter skin tones dominating the dataset (59%).

Bias can arise when the dataset does not adequately reflect the diversity of the population for which the model is intended. We can clearly see that the dataset is imbalanced and under representing certain

groups. This leads to the formation of bias. Many researchers believe that the increasing the amount of data in training the model would address the issue of bias. However, a study conducted by Ekellem and Köhler [4] demonstrates that the main issue lies in data imbalance. The study highlights that having a dataset that is both higher in quality and more diverse is more important.

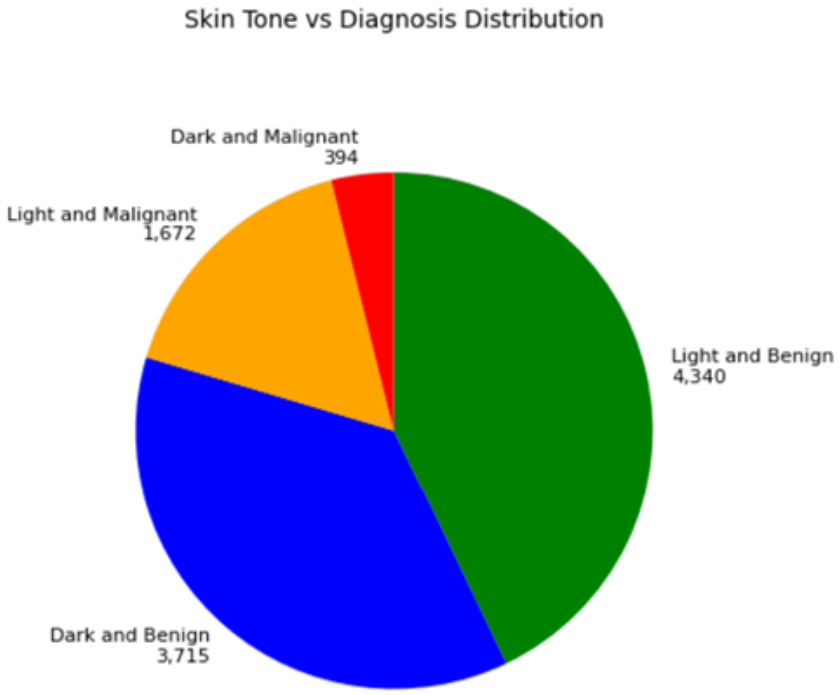

Figure 3.5: Imbalanced Dataset

3.5 Balanced Dataset

To address these issues, several strategies can be employed. Under-sampling the majority class, over-sampling the minority class (similar to bootstrapping), and generative/augmentation are few of the approaches. In this study, an approach to under-sample benign images has applied so that there are the same number of benign and malignant images. This results in much fewer images. We further under-sample light images to match dark tone images. [2]

- In the study, the under-sampling technique has been deliberately implemented to address the class imbalance which is inherently present in the original dataset, wherein benign lesions and images of lighter skin tones substantially outnumbered malignant cases and images of darker skin tones. To facilitate proper evaluation of the bias metrics, the dataset is systematically under-sampled such that each combination of skin tone (light and dark) and diagnostic category (benign and malignant) is represented by an equivalent number of samples.
- While it is acknowledged that under-sampling necessarily reduces the absolute size of the dataset and may limit the generalizability of the trained models, this trade-off is considered acceptable in the context of this fairness-oriented investigation. The primary aim is not to maximize overall classification accuracy, but to examine the bias metrics. This approach is consistent with best practices in bias and fairness research, where the construction of balanced evaluation cohorts is recognized as a prerequisite for meaningful comparative analysis. [6]

3.6 Data Annotation

In the ISIC archive, there we a lot of images which lacked the skin tone annotation. As part of the study, 150 images were taken which lacked annotation and were manually labeled based on visual inspection. Each image was classified into one of two categories—Light or Dark—depending on the appearance of

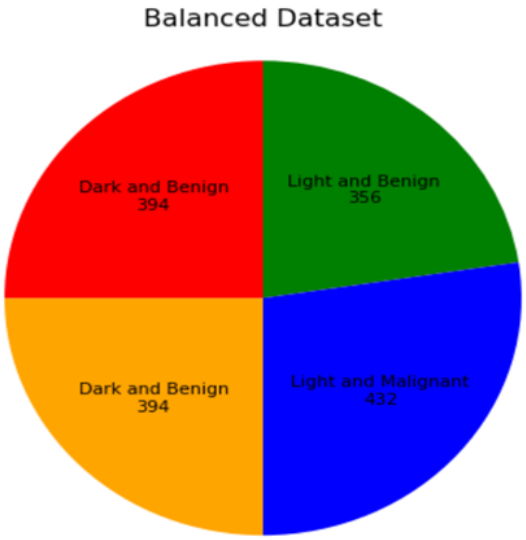

Figure 3.6: Balanced dataset

the surrounding skin in the dermoscopic photograph. These manually assigned labels were then added to the dataset metadata, enabling fairness evaluation across all images, including the extended sample.

- 150 images from the archive were labeled manually out of which only 11 had dark skin tone and rest 139 images were of light skin tone.
- Out of these 150 images, 43 were of female patients and 107 of male patients. 18 were diagnosed as malignant and 132 were benign.
- There were just two images which were of female patients with dark skin tone and diagnosed as benign. There were no images which were diagnosed as malignant in this category.
- All the 18 images which are malignant are of light skin tone. 10 of them are of female patients and 8 are of male patients.

Although manual labelling is necessary to supplement missing skin tone annotations, it introduces several limitations. Manual classification of skin tone based on visual inspection is subjective. Different annotators may interpret same images differently specially borderline cases which could be classified as either light or dark. This is why an expert intervention is required. A Non-expert labeling may overlook subtle clinical indicators of skin type, leading to misclassification. If manual labels are inconsistent or skewed toward certain tones (which is in this case), they may reinforce dataset imbalance rather than correcting it. This will lead to biased fairness evaluations, undermining the study's objective.

To overcome these limitations, annotation should ideally be performed by dermatology experts or under expert supervision.

Chapter 4

Methodology

This thesis will follow a two - stage approach where we first investigate whether the CNN based classifier models for diagnosing skin cancer are biased or not with respect to skin tone. Second, our focus will shift towards evaluating if there are any other parameter other than skin tone that can cause a bias. In the first stage, a simple image classification task is performed with the images which are ladled with skin tone and diagnosis. The objective is to determine whether it is benign or malignant. The images are categorised into two tone categories – light (Fitzpatrick Skin Type I II) and dark (Fitzpatrick Skin Type IV V). The data set used here is imbalanced with a greater number of light tone and benign images. So, to understand the imbalance another subset of the dataset will be created by sampling to have equal distribution of tones as well as malignant and benign diagnoses. Two CNN models will be implemented on both the data sets. First model being the custom CNN model which is developed based on "*SKIN CANCER MACHINE LEARNING MODEL TONE BIAS - James Pope Ayush Joshi*" and second is based on VGG16 leveraging the benifits of transfer learning. The model performance will then be evaluated to using metrices like accuracy, precision call, recall, Disparate Impact, True Positive Rate and False Positive Rate. Fairness metrics like Tone Disparate Impact and Matthew's Correlation Coefficient will be calculated to find the bias in the models. This will be followed by annotating unlabeled images to understand their impact on the results.

This approach enables a comprehensive analysis whether skin tone bias exists in the model or not.

4.1 Confusion Matrix Definition

To evaluate the fairness and potential biasness in the machine learning models, it is essential to employ some a set of quantitative bias metrics that can capture disparities. Before moving on to the fairness metrics it is important to create confusion matrix.

For a given image, a binary skin cancer classification model will predict either benign or malignant and can be compared against the true diagnosis for the image. For binary classification, the more general terms are positive and negative. We define positive to indicate malignant and define negative to indicate benign. With these definitions, we further define the following terms.[2]

- True Positive (TP): the classifier predicts malignant and the true diagnosis is malignant.
- True Positive (TP): The classifier predicts malignant and the true diagnosis is malignant.
- False Positive (FP) The classifier predicts malignant and the true diagnosis is benign.
- False Negative (FN): The classifier predicts benign and the true diagnosis is malignant.

Given a large number of images to predict, the first step in evaluating the model is to compute number of true positives, true negatives, false positives, false negatives. These four terms are usually presented in a confusion matrix with the true predictions on the diagonal and the false predictions off-diagonal.[2]

confusion matrix =

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \tag{4.1}$$

4.2 Performance Metrics

The confusion matrix serves as a fundamental tool for evaluating classification models, as it captures the counts of true positives, true negatives, false positives, and false negatives. Accuracy is one of the most widely used metrics and is defined as the proportion of correctly classified samples among all predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

While accuracy offers a summary of model performance, it can be misleading in the presence of class imbalance, as it may show poor performance for minority classes. The F1 score addresses this limitation by balancing the trade-off between precision and recall. It is defined as the harmonic mean of precision and recall values.

$$\begin{aligned} &\text{F1 score} = 2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recal}} \\ &\text{Where,} \\ &\text{Precision} = \frac{TP}{TP+FP} \text{ and Recall} = \frac{TP}{TP+FN} \end{aligned} \tag{4.3}$$

While the confusion matrix offers comprehensive insight into overall model performance, it does not inherently account for the possibility of bias in the model. To address this limitation, different **fairness metrics** are employed to quantify and compare bias between models across different clinical subgroups. This is especially important in medical machine learning models, where ethical imperatives demand not only high accuracy but also fairness and transparency in the algorithm.

4.3 Classification in CNN models

The CNN models constructed for the study, skin cancer classification is developed as a binary task, distinguishing between benign and malignant lesions. The CNN model receives dermoscopic images as input and outputs a probability distribution over the two classes via a final softmax layer. The predicted class label is assigned based on the highest output probability, with benign typically represented as class 0 and malignant as class 1. The resulting predictions are then compared to the true labels to populate the confusion matrix. This process is repeated for the entire validation set. Fig shows sample predictions.

Figure 4.1: Sample Predictions

4.4 Fairness Metrics

1. **Tone Disparate Impact:** It measures the relative rate of positive prediction between two groups. In our study, it will be the measure of positively predicting malignant cases in dark and light tones. It is represented as the following equation:

$$TDI = \frac{\text{classifier predicts cancer for dark tones}}{\text{classifier predicts cancer for light tones}} \tag{4.4}$$

2. **Sensitivity:** Sensitivity measures the proportion of True Positive cases which is correctly identified malignant cases in this study. **TP:** True Positives (malignant correctly predicted) **FN:** False Negatives (malignant missed) A lower sensitivity for one group implies a higher chance of missed cancer diagnoses, which directly translates to clinical risk and bias.

$$\text{Sensitivity } = \frac{TP}{TP + FN} \tag{4.5}$$

3. **Specificity:** Specificity measures the proportion of True negative cases which is benign cases in this study. TN: True Negatives (benign correctly predicted) **FP:** False Positives (benign misclassified as malignant) Low specificity in one subgroup may indicate over-diagnosis bias, leading to unnecessary biopsies or interventions.

$$\text{Specificity } = \frac{TP}{TP + FN} \tag{4.6}$$

4. **Area Under the ROC-AUC by Group:** The ROC curve plots True Positive Rate vs. False Positive Rate at varying thresholds. The Area Under the Curve (AUC) summarizes the model's ability to discriminate malignant from benign cases. An AUC of 0.5 indicates random guessing, while values closer to 1.0 indicate near-perfect discrimination. Comparing AUCs between groups reveals whether one subgroup consistently receives less reliable predictions.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR}) \tag{4.7}$$

5. **Predictive Quality Disparity:** It quantifies differences in the correctness of positive predictions i.e., the precision or positive predictive value between demographic groups. In the context of skin cancer classification, this metric assesses whether the proportion of images predicted as malignant that are actually malignant is equitable across both the skin tone groups.

$$\begin{aligned} &\text{Predictive Quality Disparity } = \text{ Precision }_{\text{light}} - \text{ Precision }_{\text{dark}} \\ &\text{Where,} \\ &\text{Precision } = \tfrac{TP}{TP+FP} \end{aligned} \tag{4.8}$$

6. **Matthew's Correlation Coefficient:** Unlike simple accuracy metrics, Matthew's correlation coefficient (MCC) considers the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions, providing a more comprehensive assessment of model performance, particularly in imbalanced datasets. It is a suitable metric for evaluating models in cases of class imbalance. MCC values range from-1 to 1, with 1 indicating perfect agreement, 0 indicating random prediction, and-1 indicating complete disagreement between predicted and true values.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.9}$$

4.5 Model Training

4.5.1 Custom CNN Architecture

The convolutional neural network (CNN) architecture implemented in this study is designed for binary image classification. This model has been developed in accordance to the work done in *"SKIN CANCER MACHINE LEARNING MODEL TONE BIAS – James Pope, Ayush Joshi"*[2]. The input to the very first convolutional layer is the input image. The input image is typically either a grayscale image (single channel) or a colour image (3 channels) as in this study. [7] The input to the network consists of RGB images resized to 224×224 pixels.

$$\mathrm{x} \in R^{3X224X224} \tag{4.10}$$

where, x is resized RGB image tensor. The dimension 3 corresponds to colour channel **R,G,B**.The second and third dimensions (224, 224) are the image height and width. The convolutional blocks consist of a sequential **convolutional layer (Conv2d)**, **rectified linear unit activation (ReLU)**, and **maxpoolinglayer (MaxPool2d)**.[2]

It is common to periodically insert a Pooling layer in-between successive convolution layer in a CNN architecture. Its function is to progressively reduce the spatial size of the representation to reduce the number of parameters and computation in the network, and hence to also control overfitting.[8] The architecture begins with a convolutional layer containing **32 filters** with a **7×7 kernel,** followed by two additional convolutional layers with 64 and 128 filters respectively, each using a 3×3 kernel and ReLU activation. Each convolutional layer is followed by a max-pooling layer to reduce spatial dimensionality. Pooling layers are often used within image classification tasks to reduce training times and make models more robust to noise within the learned features. These three blocks enable the model to learn hierarchical features, from low-level textures and edges in the early layers to more abstract patterns relevant for lesion classification in deeper layers. The three-dimensional output of the final convolutional block is flattened into a one-dimensional vector to transition from spatial to dense (fully connected) layers.

Dropout rate is **0.50** and the tuning process selected is the **Adam optimiser** with a **learning rate of 0.00001**.

The final classification layer is a dense layer with softmax activation, producing probabilities for the two target classes being benign (0) or malignant (1). Figure 4.2 is the model diagram of the custom CNN architecture. (Figure 4.3)

4.5.2 VGG-16 Based Transfer Learning Model

The basis of the proposed model lies in the integration of transfer learning principles with the AlexNet architecture. [9] The convolutional neural network (CNN) model discussed in this chapter has a hybrid architecture, integrating VGG16 convolutional backbone for feature extraction with an AlexNet-Style inspired fully connected head for classification.(Figure 4.2)

- **Backbone:** The feature extraction stage employs the VGG16 architecture, which is pretrained on ImageNet dataset. VGG16 is characterized by its deep but uniform design, consisting of sequential stacks of convolutional layers with 3X3 kernels, interspersed with max pooling layers of 2X2 to progressively reduce spatial resolution.

- **Classifier Head – AlexNet:** After the feature extraction, the classifier head adopts the fully connected layer which is inspired by AlexNet. It has two fully connected layers with 4096 neurons and ReLu activation layer. Dropout regularization with a **dropout rate** of **0.5** is applied to reduce overfitting. The final classification layer is a fully connected layer with 2 output units and a **softmax activation** function. 2 output units yields probabilities for the benign and malignant classes.

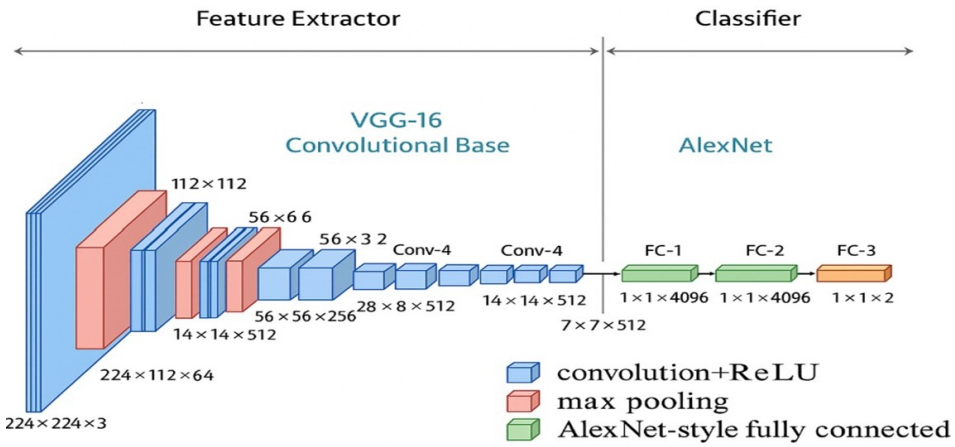

Figure 4.2: VGG-16 Based Transfer Learning Model Architecture

4.6 Experimental Setup

All the experiments were conducted using cloud based infrastructure - **Amazon Web Services (AWS)** for scalability and high performance.

- An EC2 instance - GPU instance (**g4dn.xlarge**)is provisioned as it provides **NVIDIA T4 Tensor Core GPUs** which are optimized for deep learning workloads.
- The EC2 instance is provisioned with Ubuntu 20.04 LTS, and the deep learning stack consisted of **TensorFlow 2.x**, **Keras**, **scikit-learn**, and other supporting Python libraries.
- The dermoscopic images and metadata are stored in **Amazon S3**, which served as the primary data repository.

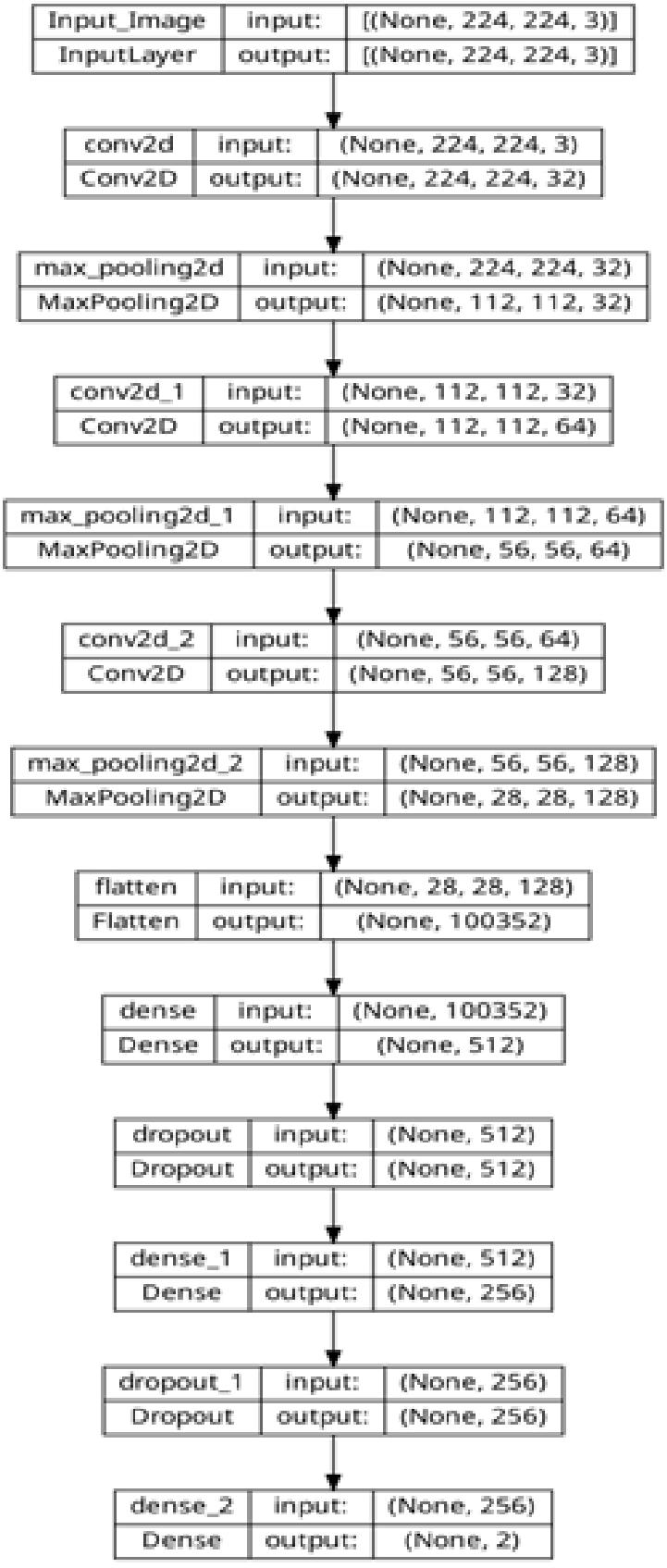

Figure 4.3: Custom CNN based Architecture

Chapter 5

Results & Discussion

5.1 Results from Custom Convolution Network

5.1.1 Results from Balanced Dataset

The first model, a custom Convolutional Neural Network (CNN) trained on a balanced dataset of skin lesion images, addresses the main focus of this thesis, which is fairness across skin tones, and achieved good diagnostic performance. With a balanced predictive ability for both benign and malignant lesions, the model reached an accuracy of 89% after 134 epochs. According to the confusion matrix, 18 of the 228 benign cases were incorrectly classified as malignant, while 210 of the cases were correctly classified. In a similar vein, 208 of the 244 malignant cases were correctly identified, while 36 were mistakenly classified as benign. According to the comprehensive classification report, benign cases were found with an F1-score of 0.89, a precision of 0.85, and a recall of 0.92. The precision, recall, and F1-score for identifying malignant lesions were 0.92, 0.85, and 0.89, respectively.

Model	**True Positive**	**True Negative**	**Total**
Positive	100	7	107
Negative	10	115	125
Total	110	122	232

Table 5.1: Confusion Matrix for Dark Skin Images

Model	**True Positive**	**True Negative**	**Total**
Positive	108	11	119
Negative	26	95	121
Total	134	106	240

Table 5.2: Confusion Matrix for Light Skin Images

Skin Tone	**Sensitivity**	**Specificity**	**ROC-AUC**	**MCC**
Light Skin Tone	*0.806*	*0.896*	*0.91*	*0.6974*
Dark Skin Tone	*0.909*	*0.943*	*0.97*	*0.8531*

Table 5.3: Comparison of Fairness Metrics by Skin Tone

$$\text{Tone Disparate Impact}_{\text{Balanced}} = \frac{\text{classifier predicts cancer for dark tone}}{\text{classifier predicts cancer for light tone}} = \frac{107/232}{119/240} = 0.930 \quad (5.1)$$

Equation (5.1) shows the TDI calculation for the balanced dataset. Table 5.1 and table 5.2 are the confusion matrix for the two groups which help in calculating the TDI.

The Tone Disparate Impact (TDI), a crucial fairness metric that measures the proportion of positive predictions made by each group, is **0.930**, which falls between the generally recognised fairness threshold of 0.8 and 1.25. Likewise, the **Predictive Quality Disparity (PQD)**, which measures the precision difference between groups, is **0.027**, which is extremely near to the optimal value of 0. Given earlier reports of tone bias in dermatological AI models, these results show a minimal difference in diagnostic predictions between light and dark skin tones, which is a noteworthy accomplishment. This conclusion was further supported by group-specific diagnostic performance. The model's **sensitivity** and **specificity** on light skin tones were **0.806** and **0.896**, respectively. With a sensitivity of **0.909** and specificity of **0.943** on dark skin tones, performance further improved.

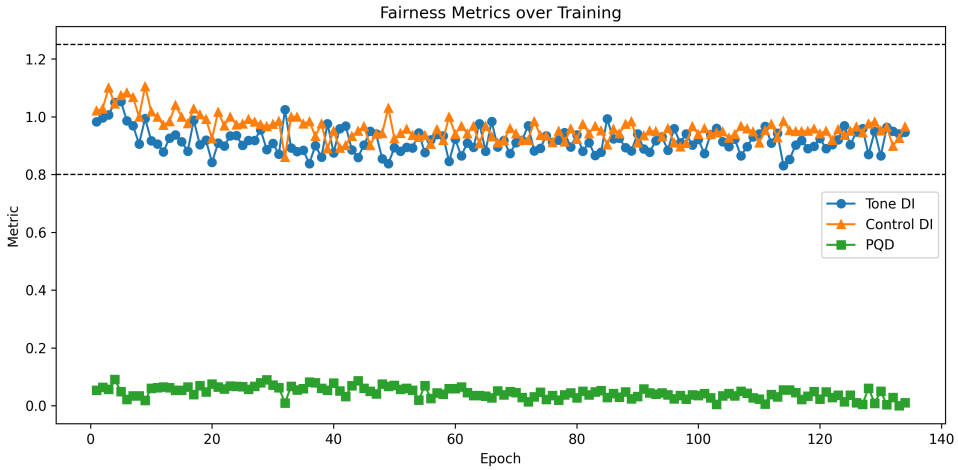

Figure 5.1: DI and PQD Curves for Balanced Dataset

To better evaluate classification reliability, **Matthews Correlation Coefficient (MCC)** and **ROC-AUC** have been computed for both groups. Light skin achieved an MCC of **0.697** and ROC-AUC of **0.910**, reflecting strong but not perfect classification quality. Dark skin, however, outperformed light skin, with an MCC of **0.853** and ROC-AUC of **0.970**, both indicative of excellent diagnostic reliability. The CDI remains consistently close to 1.0 throughout training, showing that the fairness evaluation mechanism is functioning correctly. This is expected because the control groups were artificially generated as a random split of the dataset and thus should not exhibit structural bias. This finding contrasts with prior literature, which frequently highlights poorer model performance on darker tones due to dataset imbalance.

5.1.2 Results from Imbalanced Dataset

After getting the results from the balanced dataset, the model is trained upon the imbalanced dataset. The results will help in drawing conclusion weather there is bias or not and what causes the bias in the model. model achieved an overall accuracy of 92% across the validation set, with strong performance for the benign class but reduced sensitivity for the malignant class.

- True Negatives (TN) = 2325
- False Positives (FP) = 87
- False Negatives (FN) = 141
- True Positives (TP) = 483

The results show that that predictions for benign lesions are significantly more accurate than for malignant lesions. Benign cases achieved a **precision of 0.94**, **recall of 0.96**, and **F1-score of 0.95**, whereas malignant cases lagged with a precision of 0.85, recall of 0.77, and **F1-score of 0.81**. This indicates that while the model is effective in correctly identifying benign lesions, it struggles to detect malignant lesions with the same reliability, reflecting the influence of dataset imbalance.

The confusion matrix also confirms this imbalance in prediction quality. Out of 2412 benign images, 2325 were correctly classified, while 87 were misclassified as malignant. Conversely, among 624 malignant

cases, 483 were correctly detected but 141 were missed (false negatives). This trend is particularly concerning in a clinical context, where failing to detect malignant cases can have severe consequences.

Model	**True Positive**	**True Negative**	**Total**
Positive	91	22	113
Negative	22	1061	1083
Total	113	1083	1196

Table 5.4: Confusion Matrix for Dark Skin Images

Model	**True Positive**	**True Negative**	**Total**
Positive	392	65	457
Negative	119	1264	1383
Total	511	1329	1840

Table 5.5: Confusion Matrix for Light Skin Images

Skin Tone	**Sensitivity**	**Specificity**	**ROC-AUC**	**MCC**
Light Skin Tone	*0.858*	*0.915*	*0.9481*	*0.7445*
Dark Skin Tone	*0.936*	*0.934*	*0.9790*	*0.8704*

Table 5.6: Fairness Matrix showing Sensitivity, Specificity, ROC-AUC, and MCC for different Skin Tones.

$$\text{Tone Disparate Impact}_{\text{Imbalanced}} = \frac{\text{classifier predicts cancer for dark tone}}{\text{classifier predicts cancer for light tone}} = \frac{113/1196}{457/1840} = 0.380 \tag{5.2}$$

Beyond model performance, fairness metrics provid further insight into subgroup-level disparities. **Tone Disparate Impact (TDI) is 0.380**, significantly less than the generally recognised fairness cutoff of 0.8. This highlights the bias caused by dataset imbalance by showing that patients with darker skin tones has a significantly lower chance of being predicted as positive (malignant) than those with lighter skin tones. Though malignant detection is unevenly distributed across tones, the precision remains relatively constant across groups, according to the **Predictive Quality Disparity (PQD)**, which is comparatively small at **0.052**.

Breaking down results by skin tone provided further clarity. For light skin tone images, the model achieved a **sensitivity of 0.767** and a **specificity of 0.951**, with a **Matthews Correlation Coefficient (MCC) of 0.7445** and **ROC-AUC of 0.9409**. For dark skin tone images, performance was slightly stronger: sensitivity of 0.805, specificity of 0.980, MCC of 0.7850, and ROC-AUC of 0.9790. These findings suggest that, despite the overall disparity in prediction rates highlighted by the TDI metric, the model performed marginally better at correctly classifying malignant lesions for dark-tone images when they were present in the dataset.

Fig 5.2 illustrates the evolution of fairness metrics – TDI and PQD across all the epochs when trained on imbalanced dataset. The TDI values stabilize early in training and which is significantly below the accepted range of fairness. This indicates that the **classifier predicts malignant cases for dark-skinned patients** at a much **lower rate** compared to light-skinned patients. In other words, the model exhibits a systematic bias against darker skin tones. The PQD values remain relatively close to zero across all epochs, fluctuating between 0.0 and 0.2. This suggests that the precision gap between light and dark tones is small. Once the model makes a positive prediction, its likelihood of being correct is relatively similar across tones. However, this does not prevent the bias revealed by TDI.

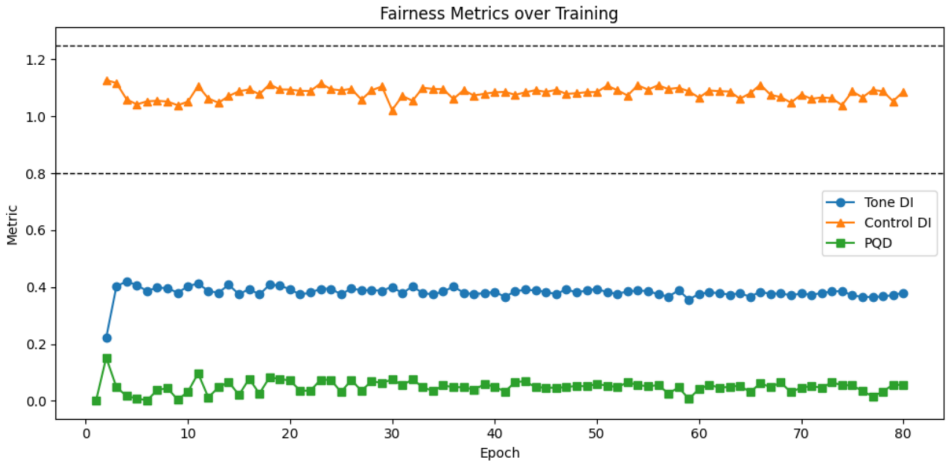

Figure 5.2: DI and PQD Curves for Imbalanced Dataset

In short, the results from the imbalanced dataset show two important things. First, while the model's overall accuracy is high, it misses a significant number of malignant cases and disproportionately favours benign classification. Second, despite group-specific sensitivity, specificity, and ROC-AUC values suggesting significant predictive ability within each subgroup, the fairness evaluation shows a bias against darker skin tones, as indicated by a low TDI score. These findings highlight how crucial dataset balancing is for reducing bias across demographic groups while improving malignant detection.

5.2 Results from VGG-16 Based Transfer Learning Model

The second CNN which is VGG-16 based transfer learning model and the results for the same are discussed below. The results have been obtained by running the model on both balanced and imbalanced data set. This will help to draw a solid conclusion if bias exists or not.

5.2.1 Results from Balanced dataset

The model achieved an accuracy of 91% with both benign and malignant cases classified with F1-scores of 0.91. Out of the 472 test images, the model correctly identified 211 benign and 218 malignant cases, with only 43 misclassifications in total.

Model	True Positive	True Negative	Total
Positive	103	8	111
Negative	7	114	121
Total	110	122	232

Table 5.7: Confusion Matrix for Dark Skin Images

Model	True Positive	True Negative	Total
Positive	115	9	124
Negative	19	97	116
Total	134	106	240

Table 5.8: Confusion Matrix for Light Skin Images

Skin Tone	Sensitivity	Specificity	ROC-AUC	MCC
Light Skin Tone	*0.767*	*0.951*	*0.9409*	*0.7445*
Dark Skin Tone	*0.805*	*0.980*	*0.9790*	*0.7850*

Table 5.9: Fairness Matrix showing Sensitivity, Specificity, ROC-AUC, and MCC for different Skin Tones.

$$\text{Tone Disparate Impact}_{\text{balanced}} = \frac{\text{classifier predicts cancer for dark tone}}{\text{classifier predicts cancer for light tone}} = \frac{111/232}{124/240} = 0.926 \quad (5.3)$$

Figure 5.3 illustrates the fairness metrics over training epochs. The Tone Disparate Impact (blue line) and Control Disparate Impact (orange line) fluctuated around 0.9–1.0, staying well within the fairness threshold bounds (0.8–1.25). This demonstrates that the model maintained stable fairness across tones throughout training. Meanwhile, the **Predictive Quality Disparity** (green line) remained consistently near zero, confirming that precision was nearly identical for light and dark skin tones.

High sensitivity for dark tones suggests that the risk of missed malignant cases or false negatives is lower, which is crucial in clinical practice. The reasonable performance across skin tones also strengthens the trustworthiness of the model which is an important requirement for deploying AI-based diagnostic tools in diverse populations. Together, these results highlight that balancing the dataset substantially mitigated bias observed in earlier experiments with imbalanced datasets. The VGG16 model not only achieved high accuracy but also delivered fair and equitable diagnostic performance across skin tones, with only minimal disparities.

Figure 5.3: DI and PQD Curves for VGG16 Balanced Dataset

5.2.2 Results from Imbalanced dataset

The VGG16 transfer learning model, trained on imbalanced dataset showed an accuracy of 92%, reflecting the strong discriminative power of the pretrained VGG16 architecture. However, the imbalance in training data led to asymmetries in performance between benign and malignant cases, as well as differences across skin tone subgroups.

	Predicted Benign	**Predicted Malignant**
Actual Benign	2369	43
Actual Malignant	196	428

Table 5.10: Overall confusion matrix for VGG16 trained on the imbalanced dataset.

Fairness Metrics: Table 5.11 shows the different parameters used in this experiment. Similar to the previous models, these fairness metrics have been calculated to understand probability of predicting malignant lesions correctly. Equation 5.4 gives the TDI = 0.447. It is a slight improvement than custom cnn model trained on the same imbalanced dataset but is way below the acceptable range highlighting a straight bias in the model.

Skin Tone	Sensitivity	Specificity	ROC-AUC	MCC
Light Skin Tone	*0.656*	*0.977*	*0.9265*	*0.7109*
Dark Skin Tone	*0.823*	*0.988*	*0.9744*	*0.8347*

Table 5.11: Fairness Matrix showing Sensitivity, Specificity, ROC-AUC, and MCC for different Skin Tones(VGG16, imbalanced datase).

	Predicted Benign	Predicted Malignant	Total
Actual Benign	1299	30	1329
Actual Malignant	176	335	511
Total	1475	365	1840

Table 5.12: Confusion matrix for light skin tone samples (VGG16, imbalanced dataset).

	Predicted Benign	Predicted Malignant	Total
Actual Benign	1070	13	1083
Actual Malignant	20	93	113
Total	1090	106	1196

Table 5.13: Confusion matrix for dark skin tone samples (VGG16, imbalanced dataset).

$$\text{Tone Disparate Impact}_{\text{imbalanced}} = \frac{\text{classifier predicts cancer for dark tone}}{\text{classifier predicts cancer for light tone}} = \frac{1083/1196}{1329/1840} = 0.447 \tag{5.4}$$

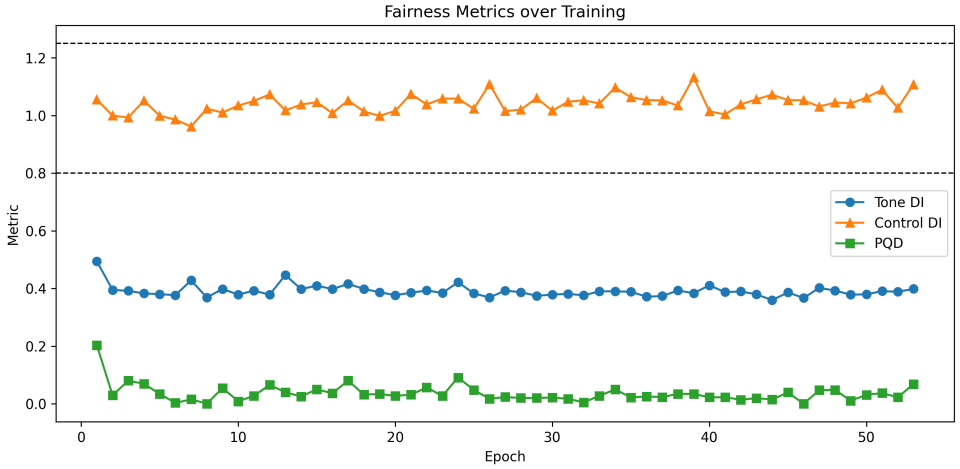

Figure 5.4: DI and PQD Curves for VGG16 Imbalanced Dataset

5.3 Per-Gender Performance Analysis

Group	n	Accuracy	F1 (Malignant)	MCC
Male	242	0.835	0.844	0.674
Female	227	0.890	0.884	0.779

Table 5.14: Per-gender performance on the diagnosis task (Malignant = 1).

Experimenting and breaking down the results by gender helps us to understand the model behaves across different subgroups. Skin-tone may not be the the only subgroup that can bring in bias. It is

Male Group	Pred 0 (Benign)	Pred 1 (Malignant)
True 0 (Benign)	94	13
True 1 (Malignant)	27	108

Table 5.15: Confusion Matrix for Male Patients (0 = Benign, 1 = Malignant).

Female Group	Pred 0 (Benign)	Pred 1 (Malignant)
True 0 (Benign)	107	12
True 1 (Malignant)	13	95

Table 5.16: Confusion Matrix for Female Patients (0 = Benign, 1 = Malignant).

very important to understand can bias arise due to some other subgroup maybe like gender. If there is disparity in the images between male and female patients then the model can bring in bias classifying malignant images. So, it becomes important to explore this domain.

- The male group achieved an accuracy of 83.5% with an F1 score of 0.844 and a Matthews Correlation Coefficient (MCC) of 0.674.
- However, the female group, performed slightly better with an accuracy of 88.9%, F1 score of 0.884, and MCC of 0.779.
- Out of 242 cases of male patients, the model misclassified 13 benign cases as malignant (false positives) and 27 malignant cases as benign (false negatives). The higher number of false negative is clinically significant because missing malignant cases can delay treatment.
- Out of 227 female cases, the model misclassified 12 benign cases and 13 malignant cases.

The gender specific evaluation indicates that though the model performed well, females achieved slightly better predictive outcomes as compared to men. This could be concluded that this difference is due to the more number of images from female patients than male.

This experiment was carried on **custom cnn model** with **balanced dataset**.

5.4 Comparative Interpretation of Results

A central finding of this study is the impact of dataset balance on both diagnostic performance and fairness outcomes. In both the model architectures investigated (Custom CNN and VGG16), the balanced dataset consistently produced more equitable results between light and dark-skinned groups, whereas the imbalanced dataset results showed clear disparities.

- When models were trained on the imbalanced dataset, malignant recall value (sensitivity) dropped substantially, particularly for light-skinned patients (0.656). This reduction meant that a higher proportion of malignant lesions in this subgroup were misclassified as benign, potentially delaying diagnosis and treatment. In contrast, dark-skinned patients achieved higher sensitivity values (0.80), suggesting comparatively better malignant detection. These differences were also reflected in the fairness metrics, with the Tone Disparate Impact (TDI) falling to 0.38–0.45, which is well below the accepted fairness threshold of 0.8.
- Training on the balanced dataset significantly improved sensitivity for both groups, reaching 0.806 (Light) and 0.909 (Dark) in the Custom CNN and 0.858 (Light) and 0.936 (Dark) in the VGG16 model. Moreover, TDI values approached parity (0.92–0.93), indicating that balancing the dataset effectively reduced disparities in prediction outcomes across skin tones.
- Incorporating manually annotated images too give a skewed result since they hold most of the images in light skin tone.

Altogether, these results demonstrate that dataset balance directly improves both classification accuracy and fairness metrics. The findings show the importance of ensuring balanced representation of diverse skin tones in dermatological training datasets.[18]

5.4.1 Clinical Implications

From the medical point of view, these findings implicate the following points:

1. The results of the unbalanced dataset showed a genuine risk: malignant cases with lighter skin tones were not properly diagnosed, which could cause treatment to be delayed. In reality, these discrepancies make already-existing healthcare disparities worse.

2. The function of diverse, balanced datasets: The enhanced fairness metrics in balanced experiments demonstrate that reducing bias can be achieved directly by selecting datasets that represent skin tones proportionately.

3. There can be more than one group that can bring in bias. Not only does skintone but also gender play a role in significantly predicting malignant or benign.

4. The process of manual annotation can be ethically challenged from the medical point of view. Either it must be done under a supervision of medical practitioner or more than one person should be involved in labeling images. Incorrect labeling of images can mislead the model and bring bias during training.

Chapter 6

Conclusion

6.1 Comparison to the previous work

The results obtained from the custom CNN in this study present both similarities and important differences when compared with the findings of "SKIN CANCER MACHINE LEARNING MODEL TONE BIAS"[2]

The concept of tone bias was introduced in the paper, showing that standard CNN architectures trained on the ISIC archive demonstrated strong overall performance but disproportionately misclassified cases in patients with darker skin tones. Their analysis revealed that the imbalance in publicly available dermoscopic datasets—dominated by lighter skin tones—was a primary source of these disparities.

The experiments conducted in this study also concludes the same. The difference stands on the results obtained. The experiment carried on balanced dataset by both the models showed better result as compared to the model. Also the results obtained by imbalanced dataset were more skewed than the previous study. The reason behind this difference could be in the dataset. ISIC is public archive and the images get added every year. The number of images used in this study is more than the images used in the previous work. As mentioned before, most of the experiments are conducted in the western countries, the images added to the dataset is also from these places. So the number of light skin images is way more than used in the previous work. Another addition to the dataset is dark skin images. In the previous work, images with FST V and FST VI were not present. Since these are newly added images, the balanced dataset formed has good number of dark skin images.

These findings suggest that the bias is not an inevitable property of CNN architectures themselves but is heavily influenced by dataset composition.

6.2 Key Takeaways from the Study

After studying both the models, and experimenting on balanced and imbalanced datasets, it can be concluded that the objective of the study has been achieved. The Aim was to study different bias metrics in the custom cnn model and develop another model for comparison. Following this analyzing if there are other subgroups which can cause bias.

- **High overall performance:** Both the custom CNN and the VGG16 transfer learning model achieved high performance in skin cancer diagnosis. This demonstrates that CNN-based approaches are highly effective in the dermatology domain.

- **Balanced Dataset Improves Fairness:** Training on a balanced dataset improves the bias as compared to imbalanced dataset. The TDI value reached close to 1 in both the models and PQD was minimal between light and dark skin tone. This highlights the importance of dataset balance in mitigating algorithmic bias.

- **Gender-Based Performance Disparities:** Extending fairness evaluation to gender revealed that the model performed better for female patients. More number of male patients had missed malignant cases which is very significant concern. This shows that bias extends beyond skin tone to other demographic subgroups.

- **Challenges of Manual Annotation:** 150 images were manually annotated using visual inspection in order to fill in the gaps in the Fitzpatrick skin tone labels in the ISIC dataset. Although expanding subgroup coverage required this, manual labeling adds inconsistency because different annotators may categories borderline cases in different ways. Furthermore, non-expert annotation runs the risk of missing subtle clinical skin tone indicators, which could reinforce bias instead of correcting it. This illustrates why tasks that are sensitive to fairness require expert-guided annotation.

6.3 Future Work

Future work on CNN models can range up to vast dimensions. Creating a robust model which can cater to all subgroups is important. As we have seen in this study, it is not just on subgroup like skin tone which can cause bias. Integrating datasets from different parts of the world will not only make the training data diverse but will also have many subgroups which can be used to mitigate the bias. We have seen that having a balanced dataset is a key to mitigate bias.

- Integrating **Generative Verifiers** in AI models can one such idea to work upon where model can be careful by making it learn to check its own mistakes and misclassifications. As we have seen, the models can misclassify and this tendency to hallucinate can be a barrier in trusting AI specially in critical fields like healthcare. With Generative verifiers, the AI learns to look at each step of its reasoning. This makes AI more careful and trustworthy.
- As building a roust model is important, deploying and running these models on suitable infrastructure is also important. GPU being a scarce commodity, it is important that the future models are run on systems which can fully maximise the GPU usage. One such usage is of **Omni** which is an internal platform built in **CLoudFlare**.[19]

Bibliography

[1] B. L. Green, A. Murphy, and E. Robinson, "Accelerating health disparities research with artificial intelligence," *Frontiers in Digital Health*, vol. 6, 2024.

[2] J. Pope, M. Hassanuzzaman, M. Sherpa, O. Emara, A. Joshi, and N. Adhikari, "Skin Cancer Machine Learning Model Tone Bias," Preprint, 2024.

[3] International Skin Imaging Collaboration Archive. [Online]. Available: https://www.isic-archive.com/. Accessed: Aug. 29, 2025.

[4] E. A. F. Ekellem and L. Köhler, "Underrepresented tones: Addressing skin bias in medical imaging for eczema, psoriasis, and melanoma detection using CNNs," in *Proc. 7th Int. Symp. Innovative Approaches in Smart Technologies (ISAS)*, Istanbul, Turkiye: IEEE, 2023, pp. 1–6, doi: 10.1109/ISAS60782.2023.10391684.

[5] J. D'Orazio, S. Jarrett, A. Amaro-Ortiz, and T. Scott, "UV radiation and the skin," *Int. J. Mol. Sci.*, vol. 14, no. 6, pp. 12222–12248, 2013, doi: 10.3390/ijms140612222.

[6] H. Suresh and J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," arXiv:1901.10002, 2019. [Online]. Available: https://arxiv.org/abs/1901.10002.

[7] Convolutional Neural Network: A Complete Guide. [Online]. Available: <insertURL>.

[8] CS231n: Deep Learning for Computer Vision. [Online]. Available: http://cs231n.stanford.edu/.

[9] A. Faghihi, M. Fathollahi, and R. Rajabi, "Diagnosis of Skin Cancer Using VGG16 and VGG19 Based Transfer Learning Models," Preprint, 2023.

[10] J. Gupta *et al.*, "Skin Cancer Classification using Deep Learning," *J. Phys.: Conf. Ser.*, vol. 2273, p. 012029, 2022.

[11] M. Shaha and M. Pawar, "Transfer Learning for Image Classification," in *Proc. 2nd Int. Conf. Electron., Commun. and Aerosp. Technol. (ICECA)*, Coimbatore, India, 2018, pp. 656–660, doi: 10.1109/ICECA.2018.8474802.

[12] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images," *Int. J. Sci. Res. Publ. (IJSRP)*, vol. 9, no. 10, p. 9420, 2019, doi: 10.29322/IJSRP.9.10.2019.p9420.

[13] T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *J. Physiol.*, vol. 195, pp. 215–243, 1968.

[14] A. W. Salehi, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," Preprint, 2021.

[15] S. Sharma, S. Sharma, and A. Anidhya, "Understanding Activation Functions in Neural Networks," *Int. J. Eng. Appl. Sci. Technol.*, vol. 4, pp. 310–316, 2020.

[16] A. Kondaka, "Evaluating Gender Bias and Fairness in Skin Lesion Diagnoses using Convolutional Neural Networks," Preprint, 2022.

[17] I. Kolm, G. Hofbauer, and R. P. Braun, "Early diagnosis of skin cancer," *Therapeutische Umschau. Revue Therapeutique*, vol. 67, no. 9, pp. 439–446, 2010, doi: 10.1024/0040-5930/a000077.

[18] H. M. Gloster and K. Neal, "Skin cancer in skin of color," *J. Am. Acad. Dermatol.*, vol. 55, no. 5, pp. 741–760, 2006, doi: 10.1016/j.jaad.2005.08.063.

[19] https://blog.cloudflare.com/how-cloudflare-runs-more-ai-models-on-fewer-gpus/?utm_source=tldrai/

Appendix A

Additional results

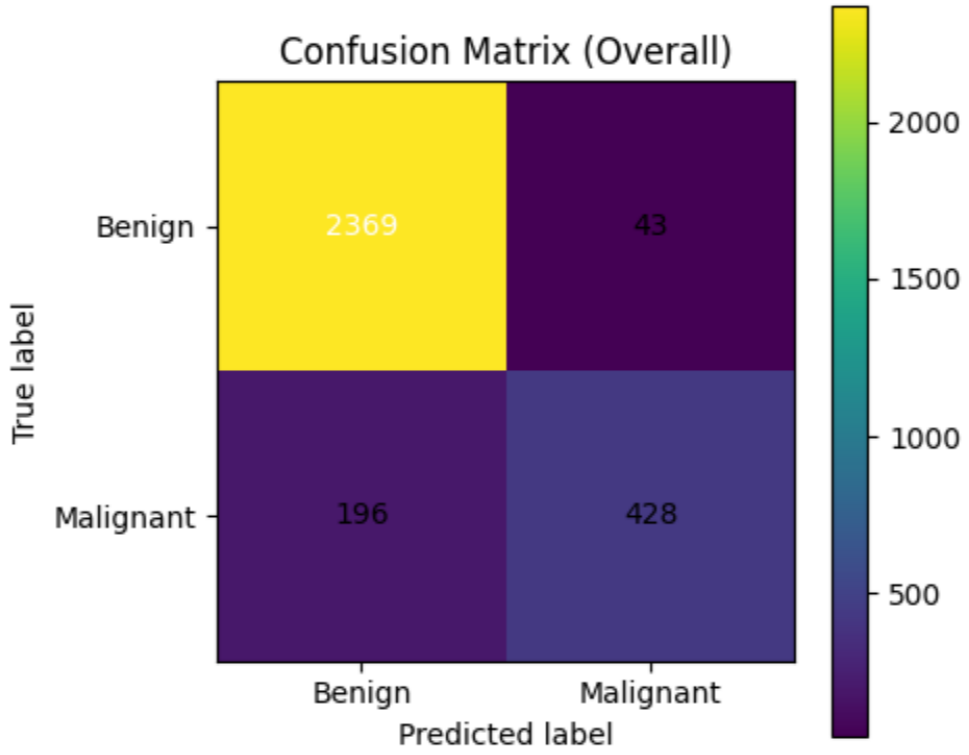

Figure A.1: confusion matrix - VGG16 model (imabalnced dataset)

Figure A.2: Group wise sensitivity and specificity - VGG16 model(imabalnced dataset)

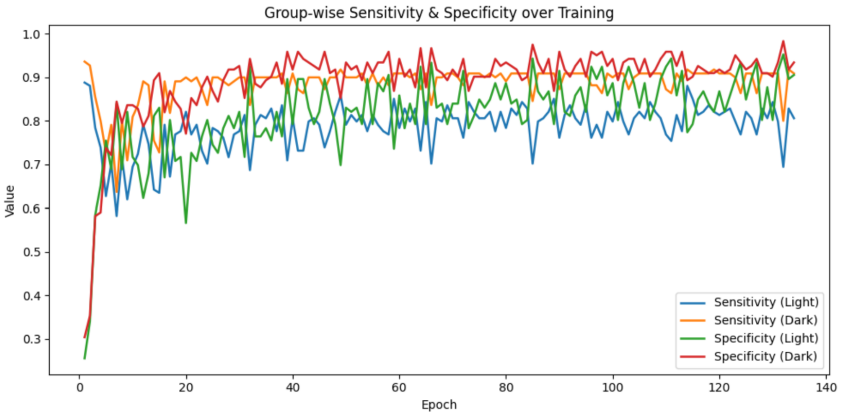

Figure A.3: Group wise sensitivity and specificity - custom cnn model(balnced dataset)

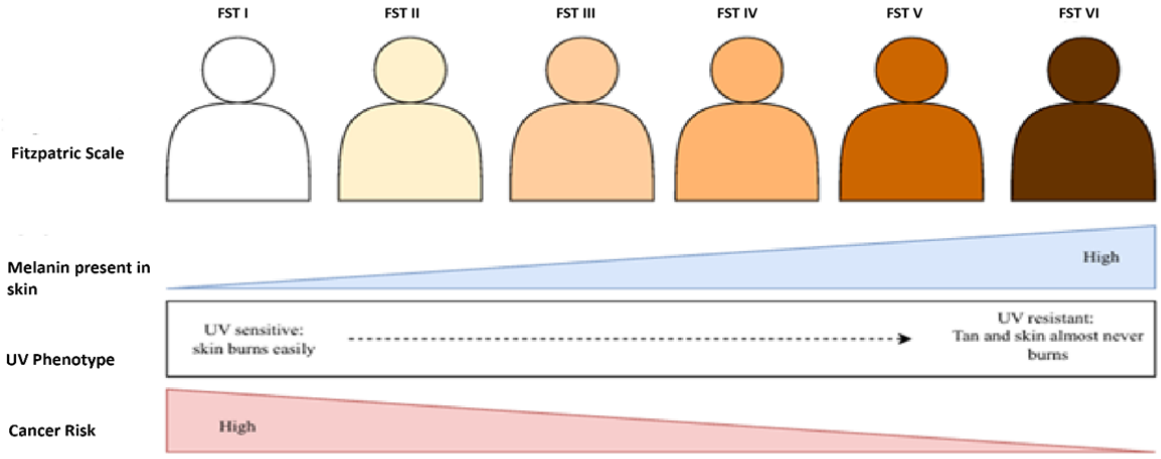

Figure A.4: Fitzpatric Scale