# Bias in Deep Learning for Skin Cancer Analysis: Challenges and Measurement Methods

University of Turku
Department of Computing
Bachelor of Science in Technology Thesis
Biomedical Engineering and Health Technology
April 2025
Sara Laanaya

UNIVERSITY OF TURKU
Department of Computing

Sara Laanaya: Bias in Deep Learning for Skin Cancer Analysis: Challenges and Measurement Methods

Bachelor of Science in Technology Thesis, 33 p.
Biomedical Engineering and Health Technology
April 2025

---

Deep learning-based AI methods are increasingly used to diagnose skin cancer lesions. However, biases in these models raise concerns about their reliability. In this context, "bias" refers to the systematic favoritism or disadvantage in the model's predictions due to unbalanced or insufficient representation of certain groups (e.g., gender and ethnicity) or features (e.g., skin color and skin lesions) in the training data. The goal of this bachelor's thesis is to evaluate different sources of bias and how they can be measured when the source of bias is known. This is accomplished by examining studies that attempt to mitigate skin color bias. The thesis provides an overview of the fundamentals of skin cancer, various imaging techniques for skin lesion analysis, the distinctions between artificial intelligence, machine learning, and deep learning, and the state-of-the-art tools specifically convolutional neural networks used for skin lesion analysis. It then explores potential sources of bias in these models and analyzes studies addressing bias mitigation, particularly regarding variations in skin tone. The studies reviewed measure bias using performance metrics that evaluate a model's ability to assess skin lesions across different skin tones, while some also calculate fairness metrics to identify bias. The final conclusions drawn from the use of fairness metrics were that, in many ways, they resemble performance metrics but offer the advantage of facilitating direct comparisons if subgroups are properly evaluated. The use of fairness metrics was often limited in the reviewed papers, partly due to the challenge of evaluating subgroups within datasets and the limitations of these metrics in assessing bias in medical diagnosis. One proposed solution is to standardize not only the images in the datasets, but also the definitions of subgroups, which would potentially enhance the usefulness of fairness metrics. Many of the current biases in deep learning models stem from a lack of diversity and insufficient data, making it difficult to design models that generalize well across all subgroups. A potential solution to this would be to develop models tailored to specific subgroups, thereby mitigating bias caused by data scarcity.

Keywords: Skin cancer, melanoma, bias, deep learning, CNN, machine learning, fairness metrics, performance metrics

TURUN YLIOPISTO
Tietotekniikan laitos

Sara Laanaya: Bias in Deep Learning for Skin Cancer Analysis: Challenges and Measurement Methods

Bachelor of Science in Technology Thesis, 33 s.
Lääketieteellinen tekniikka ja terveysteknologia
Huhtikuu 2025

---

Syväoppimiseen perustuvia tekoälymenetelmiä käytetään yhä enenevässä määrin ihosyöpien diagnosoinnissa. Malleissa esiintyvät vääristymät herättävät kuitenkin huolta menetelmien luotettavuudesta. Tässä yhteydessä "vääristymä"tarkoittaa sitä, että malli suosii tai syrjii tiettyjä ryhmiä (esim. sukupuoli, etnisyys) tai piirteitä (esim. ihonväri, ihonmuutokset) ennusteissaan, koska koulutusaineisto ei ole tarpeeksi tasapainoinen tai kattava. Tämän kandidaatintyön tavoitteena on arvioida näiden vääristymien eri lähteitä ja niiden mittaamistapoja silloin, kun vääristymän lähde on tiedossa. Kirjallisuuskatsauksessa keskitytään tutkimuksiin, joissa ihonväriin liittyviä vääristymiä pyritään vähentämään sekä esitellään mm. ihosyövän perusteet, erilaisia ihomuutosten kuvantamistekniikoita sekä tarkastellaan tekoälyn, koneoppimisen ja syväoppimisen välisiä eroja. Lisäksi käsitellään alan uusimpia ihomuutosten analysointiin käytettyjä työkaluja kuten konvoluutioneuroverkkoja. Lisäksi tarkastellaan mallien mahdollisia vääristymänlähteitä ja analysoidaan tutkimuksia, joissa pyritään minimoimaan vääristymiä (ja erityisesti ihonsävyn vaihteluun liittyen). Tarkastellut tutkimukset mittaavat vääristymiä nk. suorituskykymittareilla, joilla arvioidaan mallin kykyä analysoida ihomuutoksia eri sävyisissä ihonäytteissä. Tämän lisäksi voidaan vääristymien kvantifioimiseksi käyttää nk. oikeudenmukaisuusmittareita. Lopullisina johtopäätöksinä oikeudenmukaisuusmittareiden käytöstä todettiin, että ne muistuttavat monin tavoin suorituskykymittareita, mutta tarjoavat etuna suoran vertailun alaryhmien kesken, mikäli alaryhmät arvioidaan asianmukaisesti. Oikeudenmukaisuusmittareiden käyttö oli tarkastelluissa tutkimuksissa usein rajallista, osittain alaryhmien arvioimisen haastavuuden vuoksi, mutta myös siksi, että mittareilla on rajoituksia lääketieteelliseen diagnostiikkaan liittyvän vääristymän arvioinnissa. Yksi ehdotettu ratkaisu on standardoida paitsi aineistojen kuvat myös alaryhmien määritelmät, mikä voisi parantaa oikeudenmukaisuusmittareiden hyödyllisyyttä ja helpottaa niiden tarkastelua. Monet syväoppimismallien nykyisistä vääristymistä johtuvat monimuotoisuuden puutteesta ja riittämättömästä datasta, mikä vaikeuttaa sellaisten mallien kehittämistä, jotka yleistyisivät hyvin kaikkiin alaryhmiin. Tähän ehdotuksena oli kehittää malleja, jotka on räätälöity tiettyjä alaryhmiä varten, jolloin datan puutteesta johtuvaa vääristymää voitaisiin vähentää.

Asiasanat: Ihosyöpä, melanooma, vääristymä, bias, syväoppiminen, CNN, koneoppiminen, oikeudenmukaisuuden mittarit, suorituskykymittarit

# Contents

**1 Introduction** **1**

**2 Background** **5**

2.1 What is Skin Cancer? . . . . . 5

2.2 Automated Skin Cancer Diagnosis . . . . . 6

2.3 Understanding the Relationship Between AI, ML, and DL . . . . . 8

2.4 How Convolutional Neural Networks (CNNs) Work . . . . . 10

**3 Research Methodology** **15**

**4 Formation of Bias** **18**

4.1 Lack of Image Diversity . . . . . 18

4.2 The Functionality of a DL Model . . . . . 20

4.3 Is There No Bias If You Cannot See It? . . . . . 22

**5 Measuring Bias** **24**

5.1 Evaluating a Model's Performance Against Different Datasets . . . . 25

5.2 Fairness Metrics . . . . . 27

**6 Further Discussion** **30**

**7 Summary** **32**

**References** **34**

# 1 Introduction

Recently, there have been significant advances in AI and especially its application to medicine [1]. AI can be broadly defined as an interdisciplinary field that combines several fields such as mathematics, computer science, philosophy, economics, and neuroscience to create systems capable of solving complex problems [2]. These problems include tasks such as general problem-solving, decision-making, and pattern recognition. The newfound capability of computers to effortlessly perform tasks that demand meticulous decision-making has sparked widespread interest in their application across various medical fields, including skin cancer classification, which is the focus of this Bachelor's Thesis.

As AI develops, so do its subfields. One of the most important is machine learning (ML), which focuses on methods that learn from data. A key subset of ML is deep learning (DL), which uses hierarchical models and has become the dominant approach in tasks like image recognition (see Figure 1.1) [2]. The differences between these subfields will be discussed in Chapter 2, but in general, they can be seen as different techniques for achieving AI's goal developing software to solve complex problems across various domains.

Skin cancer is a disease of the skin that presents itself with abnormal skin lesions. These lesions are distinguishable from healthy skin and can be used in the identification processes for patient treatment [4]. DL models can analyze nonlinear relationships in distinct skin lesion patterns, thereby accelerating the classification

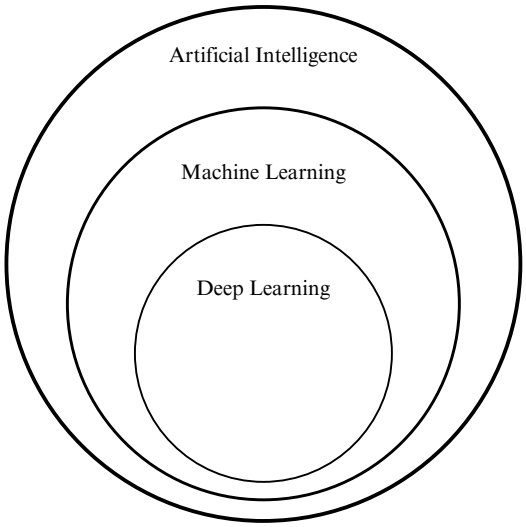



Figure 1.1: Subfields of AI [3]

process, which would typically be time-consuming if done manually by humans. Despite significant advances in the AI field, development of reliable DL models continues to be challenging. When the study of biological specimens (such as skin samples) is highly dependent on the quality and diversity of the samples, it not surprising that any type of model bias will need to be carefully considered. In other words, evaluating any systematic errors that can steer the model to make false predictions in diagnostic outcomes [5]. Acquiring diverse, high-quality medical images can also be challenging, often due to privacy regulations [6]. Luckily, for some of the challenges a solution has been introduced, for example by using "image alteration" to tackle the lack of diversity in image datasets [7]. For clarification, the term "bias" in this context refers to systematic favoritism or disadvantage in a model's predictions, often caused by unbalanced or insufficient representation of certain groups or features in the training data. However, other sources of bias, such as biases in DL model pre-training and architecture, will also be discussed in this thesis.

Despite the usage of DL in many areas, one of the major pitfalls in the development of a DL model used in the classification of skin cancer is surprisingly

enough skin color. The lack of diverse skin image datasets, leads to an imbalanced representation of skin types. A DL model's performance depends on the quality, quantity, and diversity of its training images. Most skin lesion datasets are dominated by images from patients with lighter skin, which can lead to challenges in accurately classifying skin cancer in individuals with darker skin. This bias results in higher accuracy for cancer classification in lighter-skinned patients compared to those with darker skin. [8] The early diagnosis of skin cancer is crucial for patients classified with a more severe type of skin cancer. In particular, patients diagnosed with melanoma have a 96% survival rate when detected early [9]. However, if not detected early, the survival rate drops significantly, highlighting the importance of early detection. With that being said, it is also important to note that although, people with darker skin tones have a lower risk of developing melanoma than those with lighter skin tones, a more severe type of skin cancer seems to be more predominant in people of darker skin color [10]. This emphasizes the importance of identifying and measuring bias in AI models used in the field of medicine. That way in the future, AI models can be adapted universally to save lives.

This Bachelor's Thesis sets out to find the answers to the following research questions.

- **Research Question 1:** What are the potential sources of bias in deep learning models used for skin cancer classification?
- **Research Question 2:** How is bias measured in studies that attempt to mitigate skin color bias in deep learning-based skin cancer classification?
- **Research Question 3:** What are the most promising approaches for reducing bias in deep learning models used for skin cancer classification?

The structure of this thesis will be set out in the following manner, in which Chapter 2 provides essential background knowledge, including an overview of skin

cancer classification, skin lesion imaging techniques, the key differences between AI, ML and DL, and how skin lesion images are analyzed. Chapter 3 outlines the research methodology, detailing how relevant literature was searched and selected. Chapter 4 examines how bias arises in DL models used for skin cancer classification, addressing Research Question 1. Chapter 5 evaluates methods for measuring bias, specifically in the context of skin color bias, addressing Research Question 2. Chapter 6 explores additional factors that may affect a DL model's accuracy in skin cancer classification and discusses potential solutions for reducing bias, addressing Research Question 3. And finally chapter 7 summarizes the key findings of this thesis.

# 2 Background

## 2.1 What is Skin Cancer?

The healthy human body is composed of cells. These cells follow a regular life cycle; they divide and die in a well-controlled manner. Unfortunately, cells can undergo abnormal changes that lead to uncontrolled division, a condition known as cancer. Normal skin cells may for some reason develop into cancerous skin cells, this condition is referred to as skin cancer. There are various types of skin cells, each serving a distinct role in the overall function of the skin. These skin cell types are classified with specific names, and for that reason, skin cancer can be categorized based on the type of skin cell it originates from and also how it appears when viewed under a microscope (see Figure 2.1). Skin cancer is broadly classified into melanoma and non-melanoma types. It can be categorized into multiple classes or as a binary classification. The three most prevalent types are basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma. Among these, melanoma is the most serious due to its higher likelihood of spreading to other parts of the body. One clearly fortunate aspect of skin cancer is its visibility through the formation of lesions, which often makes detection easier compared to many other types of cancer hidden within the body. [11] [4]

Skin cancer diagnosis is typically performed through a skin biopsy, where a small section of skin is removed and examined microscopically from stained samples to

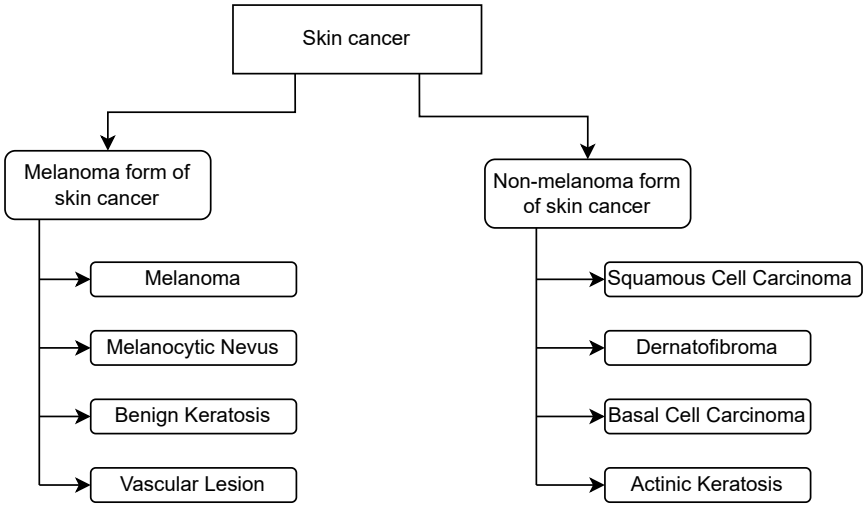



Figure 2.1: Skin cancer classification

identify cancerous cells. This approach is often necessary due to the difficulty of accurately classifying different skin cancer types by visually inspecting the lesion alone. Medical professionals also consider other factors to rule out skin cancer, such as the patient's family health history, medical conditions that might increase the risk of developing skin cancer, and changes in the appearance of skin lesions over time. [4]

## 2.2 Automated Skin Cancer Diagnosis

The most common method used by medical professionals to diagnose skin cancer is a simple visual examination of the lesion, with an accuracy of approximately 60%. When dermoscopy, a non-invasive imaging technique that magnifies and enhances skin surface structures, is used, the accuracy increases to an 89%. A dermoscopic examination is a non-invasive procedure in which a tool magnifies the lesion 10 to 20 times its original size [12]. Despite an improvement in accuracy from 60% to 89%, challenges remain, particularly in analyzing cancerous skin lesions at early stages. At these stages, lesions often lack clear, distinguishable cancerous characteristics.

Chapter 4 will further discuss these characteristics and how they are used in developing DL models and their architecture. This challenge has driven the development of computer-aided detection methods along side AI for the diagnoses of skin cancer. [13]

Several imaging modalities support the analysis and diagnosis of skin lesions, such as digital photographic imaging, confocal microscopy (CM), optical coherence tomography (OCT) and high-frequency ultrasound (HFUS). Digital photography is commonly used in dermatology due to its accessibility and low-cost. It is used to track disease progression through so-called total-body photography, however this method lacks standardization in positioning and lighting. Moreover, 3D photography devices often struggle with capturing and interpreting hair, reducing accuracy. Images taken digitally are not dermoscopic images, as dermoscopic images require the use of an external device for proper magnification. While dermoscopy is a key tool for skin cancer lesion analysis, clinicians require extensive training for its proper use. [14]

Other imaging modalities, such as reflectance confocal microscopy (RCM) and fluorescence confocal microscopy (FCM), are used to identify malignant lesions and can be performed with or without tissue removal. CM produces high-resolution black-and-white images that extend slightly beyond the first skin layer. RCM is particularly effective in diagnosing basal cell carcinomas (BCCs), while FCM, when used with tissue samples, can help identify non-melanoma skin cancers. However, interpreting CM images requires extensive training, and CM devices are expensive and their use time-consuming depending on the lesion being scanned. In contrast, OCT penetrates the skin deeper than CM and visualizes both melanoma and non-melanoma skin cancers, as well as tumor margins, but it cannot distinguish individual cells. HFUS, used to determine skin thickness and tumor depth, produces low-resolution images with limited functional contrast. [14]

As will be discussed in Chapter 4, the imaging devices used in open-access datasets are often unspecified. However, most datasets primarily feature dermoscopic images. [15] Recent advancements in AI, which excels at recognizing subtle differences in lesion features, have allowed it to surpass human performance in several cases [16]. The high cost of medical imaging equipment and the significant training required for proper use highlight the growing interest in AI models that automate skin cancer diagnosis.

Although the concept of AI emerged in 1956, the U.S. Food and Drug Administration (FDA) did not approve the first mammography computer-aided diagnosis system until 1998. Shortly after, computer-aided diagnosis was introduced in dermatology [16]. With ongoing advancements in AI, particularly in Neural Networks (NN), the development of state-of-the-art AI-driven skin cancer classification systems has continued to progress.

## 2.3 Understanding the Relationship Between AI, ML, and DL

One of the earliest methods used in the development of AI was the knowledge-based approach, which involved coding a set of formal rules that "described the world" to the computer. However, this approach proved challenging, as the formal rules needed to be carefully designed to avoid inconsistencies. [3]

These challenges led to the development of a subfield of AI: ML. Unlike the knowledge-based approach, ML trains computers to acquire knowledge by identifying patterns within raw data. In other words, ML can learn from data without relying on an extensive list of pre-coded rules. However, ML is limited by its reliance on the quality and representation of the data it extracts. [3]

To address this issue, another subfield of AI was introduced: DL. DL is a sub-

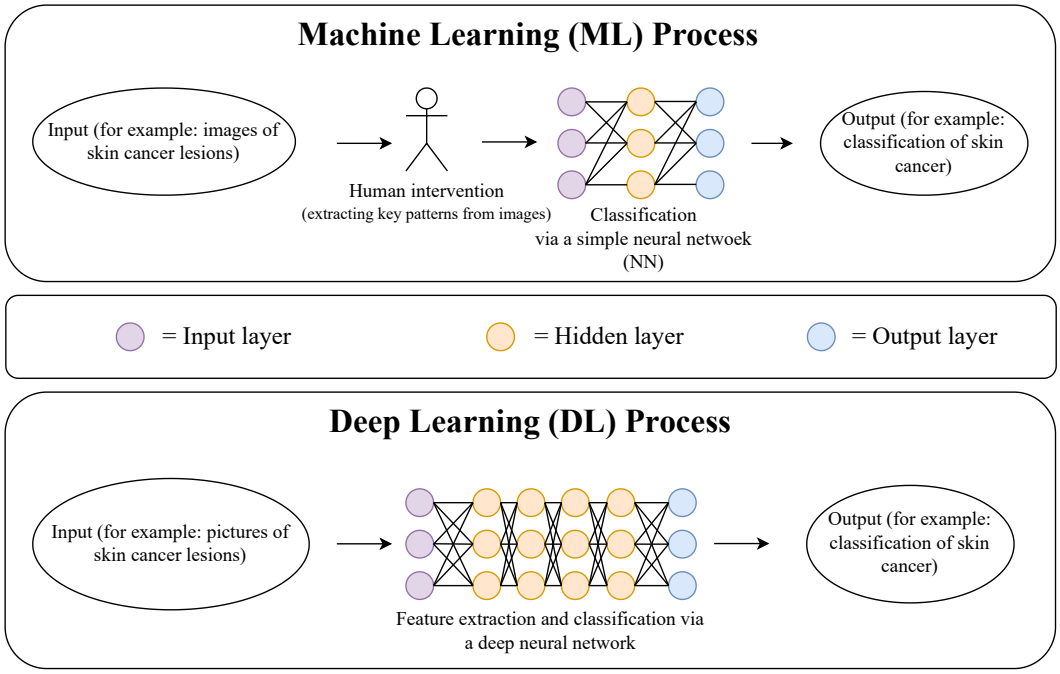



Figure 2.2: A simplified diagram of the differences between ML and DL in the classification of images.

field of ML (refer to Figure 1.1 on page 2), with its key distinction being the ability to reduce dependency on data representation by building complex concepts from simpler ones. [3] This is achieved through the use of multiple hidden layers in an artificial neural network (ANN), also called a neural network (NN), as shown in Figure 2.2. A neural network is a model designed to identify nonlinear patterns in data. The term "deep" in DL refers to the number of hidden layers in a neural network; the more hidden layers, the "deeper" the network. Deeper networks can learn more hierarchical feature representations, where lower layers capture basic patterns (e.g., edges and textures), while higher layers identify more complex structures relevant to classification. This hierarchical learning distinguishes deep networks from shallower models, which may also model non-linear relationships but lack the same depth of abstraction. [2] To illustrate the importance of this development, research by John Hopfield and Geoffrey Hinton on neural networks earned them a Nobel Prize in 2024.

[17]

The nodes, represented as yellow circles in Figure 2.2, are called neurons. The term originates from neuroscience, as brain cells are also called "neurons", though the two are not directly related. The similarity lies simply in the inspiration as the functioning of neural networks was initially modeled after the way brain cells process and transmit information. [2], [17]

The key takeaway is that AI served as the foundation for the later development of ML and DL models, with each subfield aiming to create models that are increasingly efficient and accurate while reducing time consumption. The main difference between DL and ML lies in DL models' lower dependency on data representation compared to ML models. This reduces the need for human intervention, such as manually extracting features from images. Manual extraction requires time and field experts that understand what features are important. The next subsection of this chapter will discuss how a specific DL model, the standard convolutional neural network (CNN) functions in classifying skin cancer lesions.

## 2.4 How Convolutional Neural Networks (CNNs) Work

CNNs are currently one of the most popular DL networks [18]. Understanding how a CNN works is crucial for identifying areas where bias may form and for developing methods to measure it. This thesis will only scratch the surface of how CNN models function. While it is possible to understand individual steps that a CNN takes to classify an image, the interactions of all these steps are often considered too complex to fully grasp. For this reason, DL models, including CNNs and neural networks (NNs), are often referred to as "black boxes" [19]. Although this holds some truth, it further underlines the importance of developing methods to interpret these models,

making bias detection more feasible.

Before diving into how it works, it is important to understand what an image is and how its information can be processed. An image, such as one taken with your phone, is commonly represented using the RGB color model, where it is composed of red, green, and blue pixels (see Figure 2.3). Each image varies in red, green, and blue color intensities, which can be assigned a value. Now, imagine a square-shaped image divided into individual pixels which resembles a grid. Each pixel, represented as a small square within the grid, has an assigned value, which serves as the input data for a CNN model. To extract detailed information from the image, a filter (also called a kernel or matrix) is applied. Visualize this filter as a smaller square grid that moves across the image, for all three color layers: red, blue and green. The filter has its own values, or weights, which emphasize certain features of the image for further analysis. [3], [20]

This may seem complex, as recognizing objects in daily life does not require defined steps or the weighting of specific features. For example, describing the shape of the number 8 as "two circles stacked on one another" is simple for humans. However, a computer does not automatically understand shapes like circles, which is why it requires feature extraction. In other words, a kernel facilitates this process by sliding over the image, identifying edges, and creating a new grid called a feature map [20]. The term "convolution" refers to the combination of two functions (the input and the kernel/filter/matrix) to produce a third function (the feature map) [3]. The kernel's role is to highlight essential features by assigning weights, reducing the importance of less relevant aspects like background color. For instance, in the case of the number 8, the background holds less importance than the digit itself.

As mentioned earlier, a DL model can learn and extract features from an image automatically, without the need for manual feature extraction, as required in some ML models. This automatic feature extraction is achieved by training a DL model

on correctly labeled datasets. Initially, the filter starts with randomly assigned weights and biases. These values are adjusted over time as the DL model learns from its incorrectly classified images. During this training process, the DL model moves forward from the input values to the output prediction (the "answer") refer to Figure 2.2. If the output is incorrect for instance, if the model predicts melanoma for an image when the correct label is basal cell carcinoma (BCC) the model uses a process called backpropagation. Backpropagation involves tracing back its steps to re-evaluate and adjust the weights and biases of the filters [20]. Notice that the term "bias" in the context of training a DL model refers to a parameter that allows a neuron within a neural network to adapt to features that might initially be overlooked or under emphasized. For example, due to poor lighting in an image, a neuron might receive lower pixel intensity values, resulting in a weighted sum close to zero. This could cause the neuron to treat those features as "irrelevant," even though they may carry important information. The bias value is added to the weighted sum, allowing the neuron to shift its activation threshold. As the DL model learns through training, the bias is adjusted, enabling the neuron to recognize that these values are not truly zero and ensuring that relevant features are not ignored. [20]

In addition, data often undergoes a process called preprocessing, which involves standardizing or transforming the input data (such as resizing images, normalizing pixel values, or removing noise) to ensure it is consistent and suitable for the DL model. On top of preprocessing, a process called pre-training may occur, where a model is first trained on a large, generic dataset before being fine-tuned for specific tasks, such as image classification. For instance, determining the initial weights of a kernel can be accomplished during pretraining, where the model learns general features from a large, generic dataset. [20]

Filters are essential because they reduce the amount of information fed into a CNN model. Unlike the human brain, computers have limited data-processing ca-

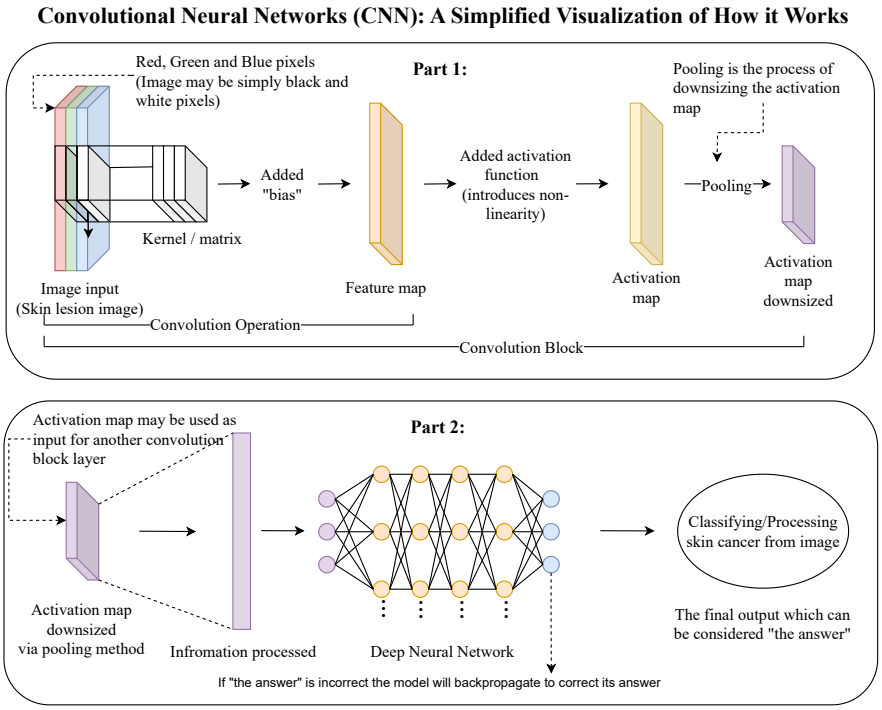



Figure 2.3: Simplified visual representation of Convolutional Neural Networks (CNN) [21] [22]

pacity, so reducing redundant or irrelevant information enhances model performance speed. However, this process could potentially be a way of introducing bias if feature extraction is not performed properly by the model.

The next stage in the convolutional layer involves applying an activation function to the output values from the feature extraction process. The purpose of an activation function is to introduce non-linearity [20]. Linear functions, such as the speed of a car accelerating at a constant rate over time, are straightforward each time interval corresponds to a uniform increase in speed. However, image data, especially in medical contexts, often requires non-linear transformations to capture the complexities of patterns, such as lesion shapes or textures.

After applying an activation function to the values obtained through convolution, an activation map is created. This activation map then undergoes a process defined as "pooling", which reduces data size while retaining the most important

features of the image. The primary goal of pooling is to make the model more efficient and robust by reducing the spatial dimensions of the feature map. It also helps mitigate overfitting, a situation where the DL model learns the training data too well, including noise and irrelevant details. Overfitting reduces the model's ability to generalize to unseen image data, resulting in lower accuracy on test or real-world data. [20]

In the final stages, the activation maps are flattened into a one-dimensional vector and passed through one or more fully connected layers. Each layer influences the next, with the final layer output used to make a prediction [20]. For example, in skin cancer classification, the final output may represent probabilities for each class, such as melanoma, basal cell carcinoma, or benign, enabling the model to make an informed diagnosis. In the following chapters, we will see that studies often simplify their CNN models to classify skin lesions in a binary manner, meaning the lesion is categorized as either melanoma or non-melanoma.

# 3 Research Methodology

To gather more information on this topic, three databases were used: PubMed, IEEE Xplore, and Web of Science. Each query has been broken down into four sections: skin-related concepts, performance-related concepts, deep learning-related concepts, and skin color-related concepts. The queries used for each database can be seen in Table 3.1. The objective is to narrow down the records to those that specifically discuss about the analysis of skin lesions using deep learning models, and how skin tone may impact its performance (this will help answer Research Question 2). The search terms for both queries (Table 3.1) were developed with the assistance of ChatGPT 4.0 and relevant keywords from references. Both search queries afterwards undergo the same refinement processes through the use of the following criteria:

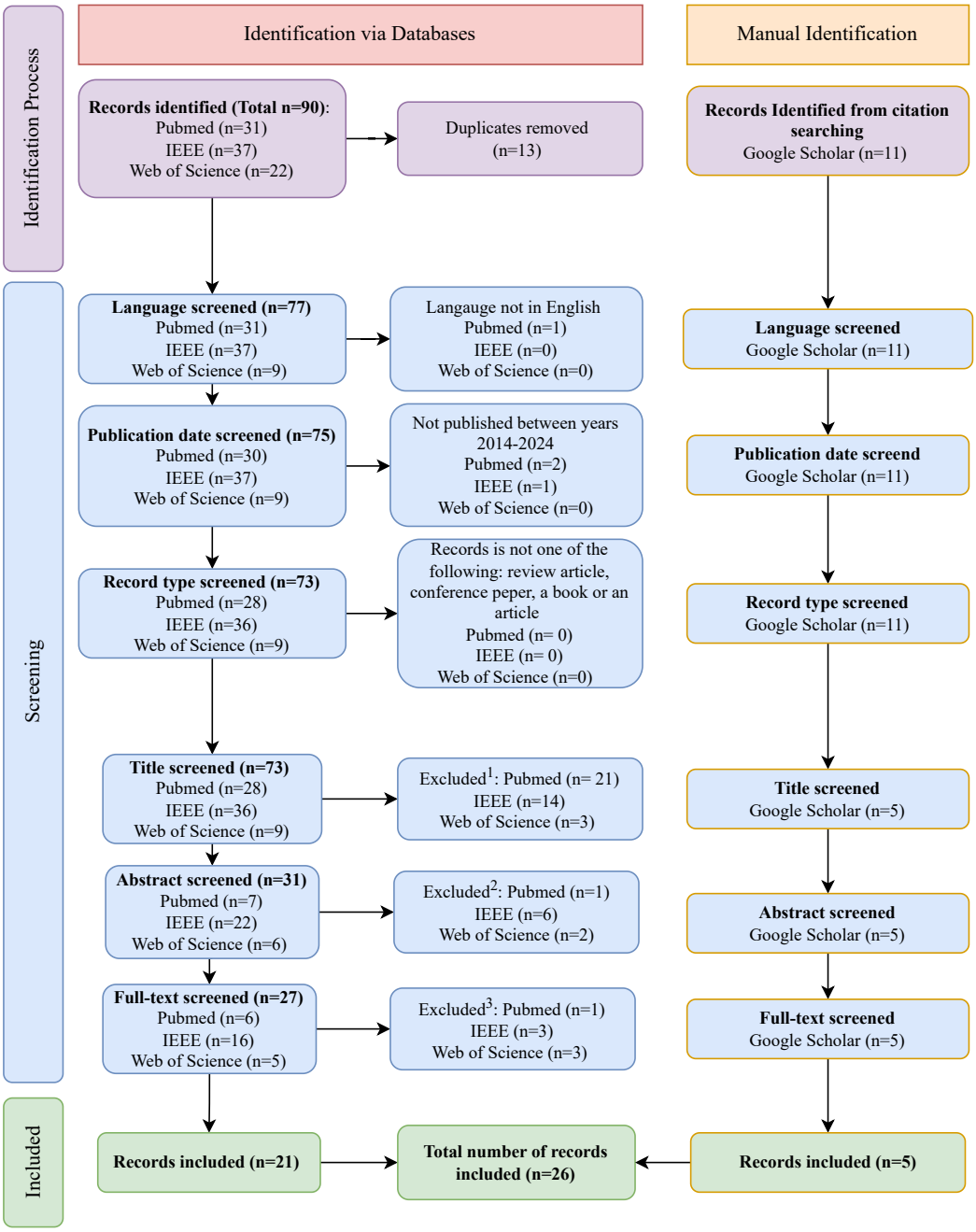
- Resources are published in English.
- Resources must be one of the following types: article, review article, conference paper, or a book.
- Resources have been published within the past 10 years (Years 2014-2024).

It is important to note that the use of review articles will only be used to provide background knowledge to readers, and that Chapters 4 and 5 will attempt to only take into consideration articles, conference papers and material written in books. During the screening process a relevant paper published in January 2024 by Benčević

Table 3.1: Search queries used in database searches

| Search Type | Query |
|---|---|
| **Specified Query for Web of Science** | ("skin cancer" OR melanoma OR "malignant melanoma" OR "skin cancer classification" OR "skin lesion" OR dermatology OR "skin lesion classification" NOT ("cosmetic dermatology" OR "plastic surgery")) AND (bias* OR "racial bias" OR debias* OR "bias-reducing methods" OR "bias mitigation techniques" OR "algorithmic bias" OR "fairness in machine learning" OR accuracy OR "model accuracy" OR "performance evaluation" OR "accuracy improvement" OR "data imbalance" OR evaluation) AND ("deep learning" OR "machine learning" OR "convolutional neural networks" OR CNN OR "classification models" OR "medical image analysis" OR "deep neural networks" OR "neural networks" OR "convolutional network") AND ("skin color" OR "skin tone variability" OR "skin tone" OR "skin shade" OR "melanin levels" OR "ethnicity" OR "skin type" OR "Fitzpatrick scale" OR "light skin" OR "dark skin" OR "skin pigmentation variability" OR "hyperpigmentation" OR "hypopigmentation" OR "skin phototype" OR "racial skin differences" OR "skin color diversity" OR "cutaneous color variation" OR "dermal pigmentation" OR "ethnic skin variations") |
| **Specified Query for PubMed and IEEE Xplore** | ("skin cancer" OR melanoma OR "malignant melanoma" OR "skin cancer classification" OR "skin lesion" OR dermatology OR "skin lesion classification") AND (bias OR "racial bias" OR debias OR "bias-reducing methods" OR "bias mitigation techniques" OR "algorithmic bias" OR fairness OR accuracy OR "model accuracy" OR "performance evaluation" OR "accuracy improvement" OR "data imbalance" OR evaluation) AND ("deep learning" OR "machine learning" OR "convolutional neural networks" OR CNN OR "classification models" OR "medical image analysis" OR "deep neural networks" OR "neural networks" OR "convolutional network") AND ("skin color" OR "skin tone variability" OR "skin tone" OR "skin shade" OR "melanin levels" OR ethnicity OR "skin type" OR "Fitzpatrick scale" OR "light skin" OR "dark skin" OR "skin pigmentation variability" OR hyperpigmentation OR hypopigmentation OR "skin phototype" OR "racial skin differences" OR "skin color diversity" OR "cutaneous color variation" OR "dermal pigmentation" OR "ethnic skin variations") NOT ("cosmetic dermatology" OR "plastic surgery") |

et al. investigated skin color bias in deep learning-based skin lesion processing, more specifically segmentation which is the process of localizing the skin lesion area of the image. This study described in great detail why certain methods used to mitigate bias did not result in a reduction of bias [23]. This paper was deemed highly suitable for this thesis, therefore, its references and related articles found through Google Scholar were reviewed manually without the use of a query. Manual search was also used to find biases outside of skin color differences which plays an important role in answering Research Question 1. After both queries (Table 3.1) undergo the same refinement process (see Figure 3.1), the resources are further processed based on their titles and abstracts. After evaluating the title and abstract it must be apparent whether or not the article discusses processes of mitigating forms of bias or evaluating sources of bias in the analysis of skin lesion images.




Exclusion criteria

[1]The title must in someway discribe bias, improvement, or fairness in a deep learning model used for skin lesions or skin cancer image analysis.

[2]The Abstract must in someway align with the object of this Bachelor's Thesis research questions.

[3]If the full text does not provide sufficient detail about model bias and how it may be mitigated in deep learning models used for the analysis of different types of skin lesions, the record will not be considered.

Figure 3.1: A PRISMA Flow Diagram

# 4 Formation of Bias

To ensure that a deep learning (DL) model functions properly, its performance must be thoroughly tested. However, testing alone is not sufficient; additional measures are needed to minimize bias and improve reliability. These include establishing proper data collection protocols, ensuring data quality control, and applying appropriate techniques during model training. This chapter explores how bias is formed, while Chapter 5 will discuss methods for measuring it. A key source of bias in DL models used for skin cancer classification is the representativeness of the training data. Bias can arise when the dataset does not adequately reflect the diversity of the population for which the model is intended. If certain subgroups are underrepresented or entirely missing, the model may perform poorly for those groups. Furthermore, decisions made during model training and pretraining, such as transfer learning strategies and other methodological choices, can introduce or amplify bias. Additionally, the architecture of the DL model itself can influence how it processes different types of input data, potentially leading to biased outcomes.

## 4.1 Lack of Image Diversity

Many researchers believe that increasing the amount of data used to train DL models would address current biases. However, a study conducted by Ekellem and Köhler [24] 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 impor-

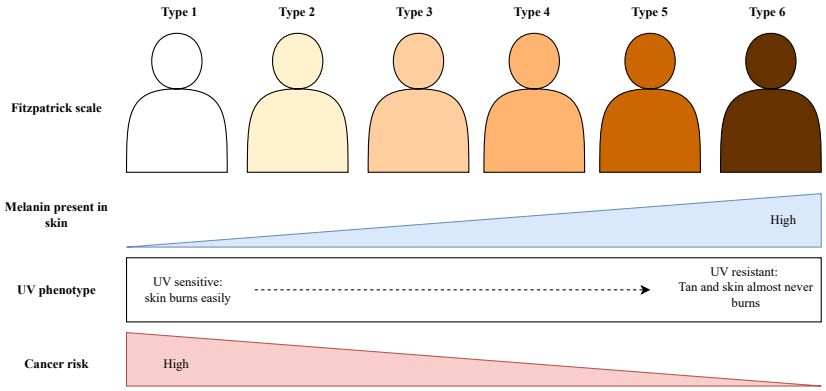



Figure 4.1: Fitzpatrick scale [26]

tant. [24] Data imbalance refers to the lack of diversity in the dataset used to develop and train a DL model. Diversity means that a dataset should adequately represent variations in skin cancer lesions which include variations in skin color types, the presence of hair, skin lesion shapes, and other image noise. It has been noted that current state-of-the-art DL models achieve high accuracy; however, they are mainly trained and evaluated on images from the same source. These image datasets often lack diversity, overrepresenting certain lesion characteristics, such as specific shapes, patterns, textures or hair, while underrepresenting others. The lack of diversity in training datasets can cause models to overfit to certain lesion characteristics, reducing accuracy for underrepresented cases.

Although skin color is easy to perceive, measuring it is challenging. The most accurate way to measure skin color is through the melanin index; however, this requires specialized equipment (reflectance spectrophotometers). Another method that has been mentioned in several of the studies was individual topology angle method. [7] The results of these measurements are often grouped into one of the following skin color types on the Fitzpatrick scale, which categorizes skin tones [25]. Figure 4.1 illustrates the Fitzpatrick scale, which can be used to assess the range of skin types that should ideally be represented and tested in the dataset [23]. A review by Wen et al. [15] reported that only 1 % of publicly available skin cancer image datasets

include information about skin color type using the Fitzpatrick scale. Additionally, Daneshjou et al. [5] found that even in datasets with Fitzpatrick scale information, other details, such as lighting conditions, were missing. These omissions can potentially lead to inaccuracies and image biases. This highlights the importance of standardized equipment, consistent conditions for evaluating skin color types, and proper training to guarantee similar results across all datasets.

## 4.2 The Functionality of a DL Model

Another potential source of bias arises during the DL model training and pretraining processes. During training, an algorithm adjusts the model's mathematical parameters based on the dataset, as described in Chapter 2. Pretraining, a form of transfer learning, involves transferring knowledge from one domain to another, enabling the model to learn faster with less data. In simpler terms, pretraining allows DL models to retain previously learned knowledge, with the fine-tuning phase further refining the model for a specific task [2]. Essentially, the more data used in pretraining, the better the model performs, much like how a student performs better on an exam when given enough practice tests. However, could different pretraining approaches unintentionally introduce bias based on how the model learns? A study by Seth, P. and Pai, A. K. [27] evaluates two commonly used pretraining methods across two different datasets, analyzing their impact on fairness and performance in a skin lesion classification model. The study found that when the pretraining dataset included more underrepresented groups, both model performance and fairness improved. However, the study also revealed differences between the pretraining approaches: when both pretraining and fine-tuning used the same distribution dataset, one method had a more limited impact on increasing fairness and performance metrics compared to the other. These findings suggest that differences in pretraining approaches must be carefully considered when designing DL models, particularly

in relation to the demographic distribution and overall diversity of the pretraining dataset.

Not all DL models process image information in the same way. As explained in Chapter 2, CNNs are commonly used for image analysis, but their reliance on spatial patterns can introduce bias. CNNs prioritize local features, which can lead to overfitting to overrepresented characteristics in the training data, making them more susceptible to bias [28]. For instance, a CNN trained on images with specific lighting conditions may struggle with cases that deviate from these patterns. In contrast, Vision Transformers (ViTs) and hybrid CNN-ViT architectures have demonstrated improved fairness and performance. Unlike CNNs, ViTs use self-attention mechanisms to analyze entire images holistically, enabling better generalization across different conditions and reducing bias. However, their higher computational and memory requirements can limit their usability on resource-constrained devices. [28]

These trade-offs emphasize the importance of carefully selecting model architectures when addressing bias in DL applications. Beyond choosing different architectures, redesigning CNN models can also enhance their performance for specific tasks. This is evident in a study by Angeline et al. [29], where the ABCD analysis method was incorporated into a CNN model to improve its ability to extract relevant information about skin lesions. The abbreviation ABCD stands for the following.

- A: Asymmetry
- B: Border
- C: Color
- D: Diameter

This detection method is used by professionals to assess if a skin lesion is a form of skin cancer or not [29]. A study found that a DL model's performance improved when ABCD features were used for feature extraction. However, the study did not

evaluate the model's performance across different skin color types, making it unclear whether the method generalizes well to diverse populations. Additionally, the ABCD detection method is based on two-dimensional features, which may introduce shape bias [30]. Since lesion contrast and border visibility can vary with skin tone, these two-dimensional properties may be more difficult to distinguish on darker skin, which should be considered when assessing the method's reliability.

## 4.3 Is There No Bias If You Cannot See It?

Bias in DL model architecture and datasets often reflects biases already present in real-world practices. For example, the lack of diversity in medical textbooks and resources depicting skin cancer lesions [7]. This would imply that, medical professionals often learn to recognize skin cancer lesions primarily from lighter-skinned patients, and this annotation bias may transfer to DL models. This lack of diversity has real-world consequences. People with darker skin tones generally have worse prognoses and lower survival rates for skin cancer compared to those with lighter skin tones [7]. Previous works have relied only on visual analysis of the skin lesion from a small group of dermatologists to provide correct skin cancer labels to images. There has been a study that showed that a model trained only on dermatologists' consensus datasets performed worse than models trained on pathologically confirmed datasets. [31] Addressing bias requires greater transparency not only in how skin lesions are analyzed but also in the availability of public data. Daneshjou et al. [5] noted that only 30% of the datasets they reviewed were accessible, making it challenging to address potential bias formation completely.

While having a diverse dataset that represents various skin color types and skin lesion types is crucial, it is equally important to consider the population and ethnicity from which the dataset is sourced. There remain unanswered questions about how and why skin cancer lesions develop and exhibit specific characteristics. It could

be possible that the shape and patterns of skin cancer lesions are influenced by ethnicity or the geographic regions where patients reside. A study by Pundir et al. [32] references research by Groh et al. [33], who investigated model accuracy imbalances across different subpopulations. Their findings indicate that the same skin disease may present differently across these groups. Pundir et al. [32] further emphasized the lack of representation for various skin diseases in datasets, indirectly highlighting the importance of including diverse skin lesion types from different subpopulations. This underrepresentation is evident in a database review by Wen et al. [15], which found that only 14 out of 21 open-access datasets reported the country of origin. Of these, 11 datasets were sourced from Europe, North America, Australia, and South Korea [15]. Due to this reason most research papers that suggest methods in ways to remove biases are often afflicted by this lack of demographic representation [34].

Chapter 5 will discuss the importance of testing models not only across a range of skin color types but also using data originating from different geographic regions. This reduces the model's tendency to overfit or overestimate performance based on limited, homogeneous datasets. Furthermore, Daneshjou et al. noted that for some databases, it was impossible to evaluate overlapping sources [5]. In some cases, groups published multiple studies but failed to clarify whether there were overlaps in data derived from the same hospital system or academic center.

In conclusion, the formation of bias in a DL model is a cumulative effect of both the data provided and the methodologies employed by the model. However, the primary source of bias concerning skin color types lies in the dataset itself, which serves as the root cause of propagating bias. Additionally, insufficient image data to capture the diverse forms of asymmetry, border characteristics, color variations, and diameters of skin cancer lesions disrupts the model's ability to learn and identify nonlinear relationships effectively.

# 5 Measuring Bias

Measuring and evaluating bias in deep learning (DL) models for skin cancer analysis is a critical step in ensuring fairness and accuracy. A study by Benčević et al. [23] highlights that many existing bias mitigation methods lack sufficient evidence due to the common practice of reporting only averaged results across all skin types. This approach obscures the individual performance variations, making it difficult to identify specific biases. For example, if one skin type produced an accuracy of 100% and darker skin tones only 60% the average would still be 80%. This chapter explores the methods used in the reviewed studies to evaluate bias, emphasizing the need for transparent and comprehensive performance reporting.

As discussed in Chapter 4, bias can arise at multiple levels, making its assessment inherently complex. Quantifying bias as a single measure is therefore nearly impossible. To quantify bias in any meaningful way, two key factors must be established: first, a clear definition of the bias, and second, an understanding of how it manifests in the model. In the study by Benčević et al. [23] the focus was on lesion segmentation evaluating and how well the model distinguishes healthy skin from cancerous lesions while comparing it to unprivileged (darker skin types) and privileged groups (lighter skin types). Overall, specific aspects of bias can be assessed using various approaches, with two commonly applied methods being:

- Evaluating model performance across different skin tones using performance metrics.

- Calculating fairness metrics to assess fairness in predictions.

Deep learning (DL) models often exhibit varying predictive performance across different skin types. Studies frequently use the Fitzpatrick Skin Type Scale to evaluate classification performance metrics, such as accuracy, precision, recall, and F1 scores, for individual skin types. These metrics provide quantitative insights into a model's ability to distinguish melanoma from non-melanoma across diverse skin tones. However, their reliability depends heavily on the composition of the dataset and the methodology used.

While this chapter primarily focuses on methods specifically designed to measure bias, it will also briefly discuss the basics of analyzing performance metrics for common binary classification tasks in DL models. A detailed examination of individual performance metrics is beyond the scope of this chapter. Following this discussion, the chapter will delve into fairness metrics and their application in evaluating bias.

## 5.1 Evaluating a Model's Performance Against Different Datasets

For multi-class classification tasks, performance metrics must be calculated separately for each class, with averages (macro, micro, or weighted) often reported. However, many studies simplify the classification problem to binary classification, distinguishing between melanoma and non-melanoma cases. In binary classification, the model defines a positive case as a skin lesion classified as melanoma and a negative case as non-melanoma. Correct predictions are labeled as "true," while incorrect predictions are labeled as "false" [29]. These predictions form the basis of a Receiver Operating Characteristic (ROC) curve, which plots Recall (the ability to correctly identify melanoma, often called "true positive rate") against the False Positive Rate (incorrectly predicting melanoma when it is not present).

The Area Under the Curve (AUC) score measures the area beneath this curve, serving as an indicator of the model's overall performance. The F1 Score combines Precision (the proportion of predicted melanomas that were correct) and Recall (the proportion of actual melanomas correctly identified) into a single value, balancing these two metrics. Many studies prefer to present ROC-AUC curves rather than just reporting accuracy, as accuracy can be misleading in imbalanced datasets (e.g., when non-melanoma cases far outnumber melanoma cases) [35]. ROC-AUC curves provide a more nuanced view of model performance and are particularly useful for identifying biases in data. They offer a better representation of how the model performs across different groups, such as those categorized by skin type or geographical location.

Here are a few case examples from the studies reviewed. Bias in medical AI is often unavoidable and difficult to identify, particularly when there is limited understanding of the medical condition and potential sources of bias. For instance, a study by Rezk et al. [7] attempted to mitigate bias by diversifying a dataset through the alteration of skin lesion images to represent darker skin tones. While the model achieved 76% accuracy, it could not be validated on real images of individuals with darker skin tones, leaving the effectiveness of the bias mitigation uncertain.

Similarly, a replication of Daneshjou et al.'s work reported comparable accuracy levels. However, the dataset used by Daneshjou et al. included a much higher proportion of non-melanoma cases than melanoma, resulting in low precision, low recall, and consequently, a poor F1 Score, as noted by J. Abhari and A. Ashok [36]. While this issue primarily reflects class imbalance rather than subgroup fairness, it illustrates how performance metrics like accuracy can be misleading. If certain subgroups are more represented in either the melanoma or non-melanoma category, class imbalance could disproportionately affect the model's ability to generalize

across those groups. Thus, although not a direct fairness issue, such imbalances can indirectly contribute to unfair outcomes if not properly accounted for.

## 5.2 Fairness Metrics

Fairness metrics are tools used to evaluate potential biases in algorithms by accounting for sensitive attributes such as age, ethnicity, gender, or, in this case, skin color [37]. A fair model should not favor or discriminate against any particular group. Among the studies found through the query, only four mentioned the use of fairness metrics to evaluate their model [27], [28], [37], [38]. Commonly used fairness metrics include those listed in Table 5.1. Disparate Impact (DI) is a ratio of positive

Table 5.1: Fairness metrics used to evaluate model bias.

| **Fairness Metric** | **Description** | **Equation** | **Variables** |
|---|---|---|---|
| **Disparate Impact (DI)** | Also called independence or demographic parity [39]. Evaluates whether decisions (e.g., melanoma detection) are distributed fairly between groups [37]. Measures the ratio of positive outcomes across groups. | $DI = \frac{P_{\text{unprivileged}}/T_{\text{unprivileged}}}{P_{\text{privileged}}/T_{\text{privileged}}}$ | $P$: Number of patients with a positive outcome. $T$: Total patients in a group. A value close to 1 indicates fairness. |
| **Predictive Quality Disparity (PQD)** | Measures whether accuracy is similar across groups, ensuring performance fairness. | $PQD = \frac{\min\left(acc_j \mid j \in S\right)}{\max\left(acc_j \mid j \in S\right)}$ | $S$: Set of skin types. $acc_j$: Accuracy for group $j$. The ratio compares the lowest and highest accuracy [28]. |
| **Equality of Opportunity Metric (EOM)** | Ensures True Positive Rate (TPR) consistency across groups [28], [37], [38]. | $EOM = \frac{1}{M}\sum_{i=1}^{M} \frac{\min\left[P(\hat{y}=1 \mid y=1, s=j),\ j \in S\right]}{\max\left[P(\hat{y}=1 \mid y=1, s=j),\ j \in S\right]}$ | $P(\hat{y}=1 \mid y=1, s=j)$: Probability of predicting a positive outcome ($\hat{y}=1$) when the correct label (the correct answer on the image in a dataset) is positive ($y=1$) for group $j$. $S$: Set of sensitive groups and M represents the total number of sensitive groups. [28] |
| **Demographic Disparity Metric (DPM)** | Measures differences in positive outcome probabilities across sensitive groups. Similar to DI, but uses probability ratios [28]. | $DPM = \frac{1}{M}\sum_{i=1}^{M} \frac{\min\left[P(\hat{y}=1 \mid s=j),\ j \in S\right]}{\max\left[P(\hat{y}=1 \mid s=j),\ j \in S\right]}$ | $M$: Total cases considered. $P(\hat{y}=1 \mid s=j)$: Probability of predicting a positive outcome for group $j$. $S$: Set of sensitive groups (e.g., skin types). Computes the ratio between min and max probabilities and averages it over $M$ [28], [39]. |

outcomes between privileged and unprivileged groups [37]. Using DI to measure has limitations. Since individuals with darker skin are statistically less likely to

develop skin cancer, their positive outcome rates are inherently lower than those of individuals with lighter skin. This discrepancy highlights how statistical fairness metrics like DI can fail to account for inherent differences between groups, such as environmental factors, lifestyle differences, and genetic factors such as the increased susceptibility of lighter skin to melanoma compared to darker skin that influence disease prevalence.

Without sufficient external data to validate the model's decisions, DI, as a metric, cannot detect bias on its own but can be used to quantify and highlight potential bias. For example, if individuals with darker skin have a lower probability of developing melanoma, false positives may lead to overdiagnosis. An analogy would be that each group is like a coin, with one group's coin being biased to land on tails (non-melanoma) more often than heads (melanoma). Without enough data, it becomes easy to either cherry-pick or misinterpret results, making biased coin appear fair. Similarly, insufficient data about a specific medical condition increases the likelihood that fairness metrics or bias assessments will fail to provide accurate results. This can be calculated by gathering enough data on disease prevalence between different subpopulations.

All studies calculated values for at least two different fairness metrics in addition to performance metrics. It is evident that performance metrics, as mentioned above, do not always provide enough information about how a model performs, especially when evaluated against sensitive groups. However, the reason that only a few studies use fairness metrics is likely because they can only be applied to models with binary classification. Additionally, all equations mentioned in Table 5.1 involve comparing positive outcomes (melanoma detected) with slight variations, focusing on accuracy, probability, or true positive rates. In other words, these fairness metrics are somewhat similar to evaluating a model's performance against underprivileged or sensitive groups, as described above. While the inclusion of fairness metrics in

the studies provided useful information in some cases, it is unclear whether they contributed significantly to detecting or mitigating biases. In certain instances, the metrics were reported as supplementary numbers without deeper analysis or application. However, in other studies, they served as a valuable tool for identifying disparities in model performance across different demographic groups. This suggests that fairness metrics can be useful when appropriately applied, but their impact depends largely on how they are interpreted and integrated into the model development process. Further exploration is needed to determine the full potential of fairness metrics in reducing bias and improving model fairness.

Fairness metrics offer a simplified view and facilitate direct comparison, which may explain why most studies do not calculate them, especially in medical diagnoses. One challenge is evaluating what subgroups should be considered within a dataset. Many studies that specifically evaluated skin color bias categorized subgroups using the Fitzpatrick scale, which was often not present in the provided datasets. This highlights the issue that many studies lack representative datasets, or even if data is available, defining relevant subgroups can be difficult. For instance, a few studies identified subgroups as either "light" or "dark" skinned, while others grouped skin color subgroups based on different Fitzpatrick scale ranges (e.g., types 1 and 2 together or types 3 and 4 together). This further emphasizes the need for standardizing not only the images in the datasets but also how subgroups are defined, to make applying and analyzing fairness metrics less challenging.

Instead of prioritizing models that generalize across large populations, a better approach may be to focus on creating models that provide personalized diagnoses for individual patients. This personalized approach could help reduce biases on a broader scale by tailoring diagnoses to specific patient characteristics.

# 6 Further Discussion

Through the analysis of CNN models in skin cancer classification, bias often arises from limited training data and an incomplete understanding of skin lesion analysis. These biases persist due to pre-existing gaps in medical knowledge, such as the underrepresentation of certain skin tones, lesion shapes, and 3D features. Increasing dataset diversity is not always the most effective solution, as its impact depends on how the deep learning (DL) model architecture processes and generalizes features. This is particularly relevant for CNNs, which are prone to overfitting. Overfitting can obscure which features the model is relying on too heavily, making it difficult to assess whether it is learning clinically relevant patterns or amplifying dataset biases. Conversely, models that generalize too well may also fail to capture subtle but important subgroup differences. Even with a diverse dataset, insufficient data can make it difficult to estimate disease prevalence within each subgroup. As a result, a model may appear biased simply because it lacks sufficient data to make reliable predictions across subgroups. Given the scarcity of representative data in certain populations, an alternative approach could be to develop models tailored to specific subgroups. This could help mitigate biases introduced by insufficient training data and improve diagnostic accuracy for underrepresented groups.

In medical contexts, fairness extends beyond equal treatment (e.g., achieving the same accuracy for all groups) to fair treatment that accounts for variations in risk and outcomes. A higher-risk population might require a more sensitive model, even

if it increases false positives, whereas a lower-risk group may prioritize specificity. Thus, some degree of bias, particularly in individual cases, can be acceptable. However, when designing AI models for large-scale use, the challenge lies in preventing harmful biases against underrepresented populations. While fairness metrics alone may not fully capture bias in medical DL models, collecting sufficiently diverse data or testing models across multiple populations and subgroups remains one of the most effective approaches. Additionally, advancements in technology are improving the ability to process larger datasets, making it increasingly likely that a more universal model for skin cancer classification will become feasible.

Addressing bias depends on a model's intended purpose, whether it is designed for a specific population or a global application. Another possible solution is to develop a general model supplemented with personalized metadata. This could help compensate for the lack of diverse image data and assist the model in making decisions, particularly for images from subpopulations it has not encountered before.

Bias is a complex challenge without a singular solution. Fairness metrics should prioritize medical relevance over technical equality, and identifying bias requires domain expertise and extensive data. Comparing populations, medical equipment, hospital settings, and lesion types could help address underrepresentation. Standardized evaluations may provide insight into key contributors to bias, though these factors likely vary across different medical conditions.

# 7 Summary

This thesis explored the importance of skin cancer awareness and the potential of artificial intelligence (AI) to support diagnosis by analyzing and classifying skin lesion images. It introduced the relationships between ML, DL, and CNNs a state-of-the-art tool in image classification. The core focus was on bias in DL models, particularly skin tone-related bias, its origins, and how studies have measured and attempted to mitigate it.

The findings showed that bias can arise from data imbalances, pretraining processes, and model architecture. Measuring bias requires a clear definition of what constitutes bias and how it affects model performance. For example, segmentation models were assessed based on performance across subgroups like different skin tones, while classification models were evaluated by comparing output ratios between underrepresented and overrepresented groups.

Although fairness metrics are commonly used in other domains, their application in medicine is limited. Disease prevalence varies across populations, making equal performance across groups difficult to define. In medical contexts, fairness metrics often resemble traditional performance metrics, which may explain their limited use in evaluating bias.

One proposed solution is developing more personalized models tailored to specific populations, rather than generalizing across all users. However, this requires

large datasets, which are difficult to obtain due to privacy regulations and underrepresentation in available data.

Bias cannot be effectively addressed without first understanding its root causes. A promising approach involves analyzing datasets and developing AI systems that can statistically evaluate lesion characteristics such as skin tone, size, shape, and texture. This could enhance model transparency and help identify which features most influence decision-making. Additionally, creating large, well-labeled, and standardized datasets with consistent image quality, clearly documented patient ethnicity, geographical information, and skin type classifications would further support fairer and more robust model development.

In conclusion, while AI holds promise for improving skin cancer detection, its success depends on addressing underlying biases. Fairness metrics alone are insufficient, especially in complex medical settings. Future efforts should aim for more representative datasets, better evaluation tools, and deeper insights into model behavior to build more equitable and clinically relevant diagnostic systems.

# References


[1] V. H. Buch, I. Ahmed, and M. Maruthappu, "Artificial intelligence in medicine: Current trends and future possibilities", *The British Journal of General Practice*, vol. 68, no. 668, pp. 143–144, 2018. DOI: 10.3399/bjgp18X695213.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. England: Pearson Education, 2016, ISBN: 9781292153964.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. England: MIT Press, 2016, ISBN: 9780262035613.

[4] K. Agnew, B. Glichrest, and C. Bunker, *Fast Facts: Skin Cancer - Clinical Features and Diagnosis*. Abingdon: Health Press Limited, 2005, ISBN: 9781908541390, 9781908541536, 9781908541505.

[5] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review", *JAMA Dermatology*, vol. 157, no. 11, pp. 1362–1369, Nov. 2021. DOI: 10.1001/jamadermatol.2021.3129.

[6] S. Padmapriya and S. Parthasarathy, "Ethical data collection for medical image analysis: A structured approach", *ABR*, vol. 16, pp. 95–108, 2024. DOI: 10.1007/s41649-023-00250-9.

[7] E. Rezk, M. Eltorki, and W. El-Dakhakhni, "Leveraging artificial intelligence to improve the diversity of dermatological skin color pathology: Protocol for

an algorithm development and validation study", *JMIR Research Protocols*, vol. 11, no. 3, e34896, 2022. DOI: 10.2196/34896.

[8] E. Rezk, M. Eltorki, and W. El-Dakhakhni, "Improving skin color diversity in cancer detection: Deep learning approach", *JMIR Dermatol*, vol. 5, no. 3, e39143, Aug. 2022. DOI: 10.2196/39143.

[9] B. Khatri, "Skin cancer detection: A survey", *International Journal of Research in Science and Technology*, vol. 13, pp. 01–03, Jan. 2023. DOI: 10.37648/ijrst.v13i01.001.

[10] L. Barros, L. Chaves, and S. Avila, "Assessing the generalizability of deep neural networks-based models for black skin lesions", in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, V. Vasconcelos, I. Domingues, and S. Paredes, Eds., Cham: Springer Nature Switzerland, 2024, pp. 1–14. DOI: 10.1007/978-3-031-49249-5_1.

[11] T. Slevin, Ed., *Sun, Skin and Health*. Australia: CSIRO Publishing, 2015, ISBN: 9781486301164.

[12] D. Surendren and J. Sumitha, "Machine learning algorithms for skin cancer diagnosis: Comparative analysis", in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India: IEEE, 2023, pp. 608–613. DOI: 10.1109/ICIRCA57980.2023.10220845.

[13] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H.-C. Kim, "Skin cancer detection using deep learning—a review", *Diagnostics (Basel)*, vol. 13, no. 11, p. 1911, 2023. DOI: 10.3390/diagnostics13111911.

[14] S. L. Schneider, R. Kornik, C. A. Egan, *et al.*, "Emerging imaging technologies in dermatology: Part ii: Applications and limitations", *Journal of the American Academy of Dermatology*, vol. 80, no. 4, pp. 1121–1131, 2019. DOI: 10.1016/j.jaad.2018.11.043.

[15] D. Wen, S. M. Khan, A. Ji Xu, *et al.*, "Characteristics of publicly available skin cancer image datasets: A systematic review", *The Lancet Digital Health*, vol. 4, no. 1, e64–e74, Jan. 2022. DOI: 10.1016/S2589-7500(21)00252-1.

[16] Z. Li, K. C. Koban, T. L. Schenck, R. E. Giunta, Q. Li, and Y. Sun, "Artificial intelligence in dermatology image analysis: Current developments and future trends", *Journal of Clinical Medicine*, vol. 11, no. 22, p. 6826, 2022. DOI: 10.3390/jcm11226826.

[17] T. R. S. A. of Sciences, "Popular information: The nobel prize in physics 2024", trans. by C. Barnes, A. I. Ulf Danielsson Olle Eriksson and E. Moons, Eds., 2024, Text: Anna Davour, Illustrations: Johan Jarnestad, Editor: Sara Gustavsson. [Online]. Available: https://www.nobelprize.org/prizes/physics/2024/popular-information/ (visited on 12/01/2024).

[18] L. Alzubaidi, J. Zhang, A. J. Humaidi, *et al.*, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions", *Journal of Big Data*, vol. 8, no. 1, p. 53, 2021. DOI: 10.1186/s40537-021-00444-8.

[19] J. Dobson, "On reading and interpreting black box deep neural networks", *International Journal of Digital Humanities*, vol. 5, pp. 431–449, 2023. DOI: 10.1007/s42803-023-00075-w.

[20] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*. Cham: Springer International Publishing, 2023, ISBN: 9783031296420. DOI: 10.1007/978-3-031-29642-0.

[21] O. Sevli, "A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation", *Neural Computing and Applications*, vol. 33, pp. 12 039–12 050, Sep. 2021. DOI: 10.1007/s00521-021-05929-4.

[22] S. Mallick, "Understanding convolutional neural networks (CNN)", 2020. [Online]. Available: https://learnopencv.com/understanding-convolutional-neural-networks-cnn/ (visited on 03/03/2025).

[23] M. Benčević, M. Habijan, I. Galić, D. Babin, and A. Pižurica, "Understanding skin color bias in deep learning-based skin lesion segmentation", *Comput Methods Programs Biomed*, vol. 245, p. 108 044, Mar. 2024. DOI: 10.1016/j.cmpb.2024.108044.

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

[25] A. Benmalek, C. Cintas, G. A. Tadesse, R. Daneshjou, K. R. Varshney, and C. Dalila, "Evaluating the impact of skin tone representation on out-of-distribution detection performance in dermatology", in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece: IEEE, 2024, pp. 1–5. DOI: 10.1109/ISBI56570.2024.10635847.

[26] J. D'Orazio, S. Jarrett, A. Amaro-Ortiz, and T. Scott, "UV radiation and the skin", *International Journal of Molecular Sciences*, vol. 14, no. 6, pp. 12 222–12 248, 2013. DOI: 10.3390/ijms140612222.

[27] P. Seth and A. K. Pai, "Does the fairness of your pre-training hold up? examining the influence of pre-training techniques on skin tone bias in skin lesion classification", in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA: IEEE, 2024, pp. 580–587. DOI: 10.1109/WACVW60836.2024.00067.

[28] C. Yang, Y. Sheng, P. Dong, *et al.*, "Fast and fair medical AI on the edge through neural architecture search for hybrid vision models", in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, San Francisco: IEEE, 2023, pp. 01–09. DOI: 10.1109/ICCAD57390.2023.10323652.

[29] J. Angeline, A. Siva Kailash, J. Karthikeyan, *et al.*, "Automated prediction of malignant melanoma using two-stage convolutional neural network", *Archives of Dermatological Research*, vol. 316, p. 275, 2024. DOI: 10.1007/s00403-024-03076-z.

[30] A. Lucieri, F. Schmeisser, C. P. Balada, S. A. Siddiqui, A. Dengel, and S. Ahmed, "Revisiting the shape-bias of deep learning for dermoscopic skin lesion classification", in *Medical Image Understanding and Analysis*, G. Yang, A. Aviles-Rivero, M. Roberts, and C.-B. Schönlieb, Eds., ser. Lecture Notes in Computer Science, vol. 13413, Cham: Springer, 2022, pp. 46–61. DOI: 10.1007/978-3-031-12053-4_4.

[31] M.-C. Chiu, Y. Wang, Y.-J. Kuo, and P.-Y. Chen, "DDI-CoCo: A dataset for understanding the effect of color contrast in machine-assisted skin disease detection", in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea: IEEE, 2024, pp. 6905–6909. DOI: 10.1109/ICASSP48485.2024.10448011.

[32] A. Pundhir, S. Verma, and B. Raman, "Towards ethical dermatology: Mitigating bias in skin condition classification", in *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan: IEEE, 2024, pp. 1–8. DOI: 10.1109/IJCNN60899.2024.10650487.

[33] M. Groh, C. Harris, L. Soenksen, *et al.*, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset", in

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA: IEEE, 2021, pp. 1820–1828. DOI: 10.1109/CVPRW53098.2021.00201.

[34] R. L. Correa-Medero, B. Patel, and I. Banerjee, "Adversarial debiasing techniques towards 'fair' skin lesion classification", in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, Baltimore, MD, USA: IEEE, 2023, pp. 1–4. DOI: 10.1109/NER52421.2023.10123788.

[35] S. Yan, Z. Yu, X. Zhang, *et al.*, "Towards trustable skin cancer diagnosis via rewriting model's decision", in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, 2023, pp. 11 568–11 577. DOI: 10.1109/CVPR52729.2023.01113.

[36] J. Abhari and A. Ashok, "Mitigating racial biases for machine learning based skin cancer detection", in *Proceedings of the Twenty-Fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, New York, NY, USA: Association for Computing Machinery, 2023, pp. 556–561, ISBN: 9781450399265. DOI: 10.1145/3565287.3617639.

[37] T. Oguguo, G. Zamzmi, S. Rajaraman, F. Yang, Z. Xue, and S. Antani, "A comparative study of fairness in medical machine learning", in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, Cartagena, Colombia: IEEE, 2023, pp. 1–5. DOI: 10.1109/ISBI53787.2023.10230368.

[38] J. Wang, Y. Zhang, Z. Ding, and J. Hamm, "Achieving reliable and fair skin lesion diagnosis via unsupervised domain adaptation", in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2024, pp. 5157–5166. DOI: 10.1109/CVPRW63382.2024.00523.

[39] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023, ISBN: 9780262048613. [Online]. Available: `https://fairmlbook.org/`.