School of Computer Science and Information Technology
Lucerne University of Applied Sciences and Arts (Switzerland)

Demographic Biases in Dermatology Models

Bachelor Thesis

presented to School of Computer Science and Information Technology of Lucerne University of Applied Sciences and Arts (Switzerland) in consideration for the award of the academic grade of *Bachelor in Computer Science.*

by

Nadja Stadelmann

from

Lucerne (Switzerland)

Declaration

Bachelor Thesis at Lucerne University of Applied Sciences and Arts
School of Computer Science and Information Technology

Title of Bachelor Thesis:	Demographic Biases inDermatology Models
Name of Student:	Nadja Stadelmann
Degree Program:	Bachelor in Computer Science
Year of Graduation:	2025
Main Advisor:	Dr. Ludovic Amruthalingam
External Expert:	Dr. Jürg Schelldorfer
Industry partner/provider:	Applied AI Research Lab

Code/Thesis Classification

☒ Public (Standard)
□ Private

Declaration

I hereby declare that I have completed this thesis alone and without any unauthorized or external help. I further declare that all the sources, references, literature and any other associated resources have been correctly and appropriately cited and referenced. The confidentiality of the project provider (industry partner) as well as the intellectual property rights of the Lucerne University of Applied Sciences and Arts have been fully and entirely respected in completion of this thesis.

Place/Date, Signature ________________________

Submission of the Thesis to the Portfolio Database

Confirmation by the student

I hereby confirm that this bachelor thesis has been correctly uploaded to the portfolio database in line with the code of practice of the University. I rescind all responsibility and authorization after upload so that no changes or amendments to the document may be undertaken.

Place/Date, Signature ________________________

Expression of thanks and gratitude

Thanks to my family, relatives and friends for all the support given to finish this thesis. TODO: add thanks and gratitude Ludovic Amruthalingam Simone Lionetti - deputy Ludovic Pascal Baumann - LaTeX Philippe Gottfrois - information and work on PASSION project Proofreaders TODO: do you want to be mentioned with name or not?

Nadja Stadelmann, 2025

Intellectual property of the degree programs of the Lucerne University of Applied Sciences and Arts, FH Zentralschweiz, in accordance with Student Regulations: Studienordnung

Summary

TODO: Your abstract here. The content of your thesis in brief.

Contents

Todo list

TODO: also solve todos in the code ;)

TODO: remove all \rawcitationstart \rawcitationend \baaCriteria

TODO: TEXT MISTAKES ensure fine-tuning ResNet-50, overview of instead of overview over, accuracy (you got all other versions of rr and cc); coma after e.g., point after vs., decision not desicion, decision-making not decision making ...

TODO: when you got to many pages: fix in order to - to, for the purpose of - For

Alle Fakten (fundiertes Wissen Dritter) sind korrekt zitiert. Es werden verschiedene Zitierweisen verwendet und teilweise mehrere Interpretationen gegenübergestellt. Der gemeinsam definierte Zitierstil im Text, in Abbildungen und Tabellen sowie im Literaturverzeichnis wird korrekt und durchgängig angewendet. Eigene Leistungen (sowie Bewertungen) und Fremdquellen sowie Recherchen sind klar unterscheidbar.

Die erstellten Artefakte sind von sehr hoher Qualität. Das trifft u.a. auf Diagramme, Skizzen sowie Notationen (z.B. BPMN/UML) zu. Darstellungen sind einwandfrei, alle statistisch notwendigen Qualitätskriterien sind erfüllt. Beschriftungen etc. sind vorhanden, keine Einwände, Text und Bild stimmen beschreibend gut überein. Es wurden angemessene Dokumentationsmethoden und -arten korrekt verwendet. Vereinbarte Interview Transkripte, Beobachtungsprotokolle bzw. Zusammen-fassungen sind vorhanden. Daten, Ort, Kontext, Beschreibung, Zeilennummer, Verweise, Strukturen sind erkennbar, gut formatiert und korrekt mit dem Text/ der Analyse verknüpft. Alle Elemente und Themen sind im methodischen Teil/Text erklärt und verständlich, keine technischen oder strukturellen Einwände. Auch Zwischenanalysen, Zwischenschritte oder Gesamtauswertungen wurden durchgeführt, die Herkunft der Daten ist erkennbar und professionell aufbereitet.

Der Schreibstil aller Dokumente entspricht hohen Standards und enthält keine Übertreibungen oder unbegründete Beurteilungen. Die Sprache ist aussagekräftig, prägnant und präzise. Die Fachterminologie ist konsistent, d.h. für gleiche Gegenstände und Themen werden immer die gleichen Begriffe verwendet. Der Sprachgebrauch ist durchgängig geschlechtergerecht, einheitlich und sachlich.

List of Figures

List of Tables

Glossary

equalized odds difference The absolute difference in true positive and false positive rates between subgroups, used as a group fairness metric.. xi, 38,

equalized odds ratio The ratio of true positive and false positive rates between subgroups, used as a group fairness metric.. xi, 38,

Fairlearn A Python library for assessing and improving fairness in machine learning models. It supports various fairness metrics and mitigation techniques, especially for binary classification tasks (contributors, n.d.).. 11, 33, 34, 37, 38, 41, 42, 49, 60, 61

Fitzpatrick skin type A skin classifier based on the skins' reaction to ultraviolet light, developed by dermatologist Dr. Thomas Fitzpatrick (Gottfrois et al., 2024). xi, 4,

GPUhub Lucerne University of Applied Sciences and Arts (HSLU)'s server infrastructure for GPU-related computing. It provides isolated environments with JupyterLab access for developing and running Machine Learning (ML) workflows.. 31, 33

Jupyter Notebook Executable files, often used in ML to write Python code and add explanations in text form. 5, 40, 70

pediatric A medical term for infants, children and adolescents. 1, 4, 46

proxy variable "one or more variables that encode the protected attribute with a substantial degree of accuracy" according to https://medium.com/bcggamma/practice-ai-responsibly-with-proxy-variable-detection-42c2156ad986. 29, 32, 39, 44, 45, 103

teledermatology dermatological care from a distance, supported by modern technology (Pala et al., 2020). 1, 4, 28, 31, 32, 110

Acronyms

AI Artificial Intelligence. 1, 3, 7, 8, 10, 12, 19, 20, 23, 24, 25, 26, 28, 31, 32, 95, 96, 97, 107

EOD equalized odds difference. *Glossary:* equalized odds difference, 38, 39, 41, 43, 51

EOR equalized odds ratio. *Glossary:* equalized odds ratio, 38, 39, 41, 43, 56

FPR false positive rate. 10, 42, 43

FST Fitzpatrick skin type. *Glossary:* Fitzpatrick skin type, viii, 4, 5, 26, 33, 38, 41, 44, 45, 46, 48, 50, 55, 56, 70, 100

HSLU Lucerne University of Applied Sciences and Arts. x, 31

ML Machine Learning. vii, x, 3, 4, 7, 8, 9, 10, 13, 14, 15, 16, 17, 18, 26, 55, 56, 97, 98, 99, 101, 107, 120

TPR true positive rate. 10, 42, 43, 48, 121

TODO: fix citations in glossary

1 Problem Statement

Welche Ziele, Fragestellungen werden mit dem Projekt verfolgt? Die Bedeutung, Auswirkung und Relevanz dieses Projektes für die unterschiedlichen Beteiligten soll aufgeführt werden. Typischerweise wird hier ein Verweis auf die Aufgabenstellung im Anhang gemacht.

In Sub-Saharan Africa dermatology treatment is inaccessible according to Gottfrois et al. (2024). There is fewer than one dermatologist available per one million people. Despite this, up to 80% of the children and adolescents in the area are affected by skin conditions. Teledermatology based on Artificial Intelligence (AI) promises to close this gap of specialists per case, for example by serving as a triage option. Potential patients could upload pictures to diagnostic dermatology AIs which can indicate whether the person should indeed visit a dermatologist or promote other treatment options. However, current dermatology AIs tend to fail to deliver accurate results for patients with highly pigmented skin tones. This is mainly due to demographic biases in existing AI models. The models are trained on established datasets which mainly feature low pigmented skin. Therefore, the datasets lack representation of highly pigmented skin, leading to AI models which do not generalize to the population in Sub-Saharan Africa (Gottfrois et al., 2024).

These biases result in unequal access to treatment and especially affect underrepresented groups. Such biased results must be avoided, especially in AI models which impact life-changing decisions (Mehrabi et al., 2021).

According to Diaz et al. (2022), demographic biases are especially important in dermatology. Demographic differences in patients influence the appearance of dermatological conditions. The differences in appearance can be developed depending on genetic factors, such as skin tone, age and sex (Diaz et al., 2022). Research showed, that in patients with lower socio-economic status the disease progression is more advanced at time of diagnosis, which in turn can lead to different appearances for the same disease (British Association of Dermatologists (BAD), 2021). Since the AI models use pictures as the inputs and can only learn to diagnose diseases according to their appearances in the data, the factors which affects the disease appearances must be considered when creating an inclusive dataset.

In order to overcome these issues, the PASSION research team founded the PASSION project. The projects vision is to make dermatology treatment accessible in Africa by enabling the AI-supported teledermatology for triage by reducing the demographic biases in the dermatology AI models. For this bias mitigation, the researcher collected a dataset in Sub-Saharan Africa, focusing on patients with highly pigmented skin and the most common regional pediatric skin conditions.

The PASSION dataset is complementary to existing datasets and improves their diversity. With this dataset, the PASSION team trained a ResNet-50 model which was pretrained on ImageNet. This thesis refers to this trained model as the PASSION model. It should serve as a benchmark model to assess other dermatology models in regards of fairness (Gottfrois et al., 2024). TODO: check sources, maybe, for the last sentence, the midterm protocol must be cited instead

So that the PASSION model can become an unbiased benchmark model, potential demographic biases in it must be reduced as far as possible. To reach this goal, demographic biases in the model as well as the limitation of the gathered dataset must be identified and mitigated. This thesis supports the PASSION team in this process. The main objective of the thesis is to assess the effectiveness of mitigation strategies to reduce demographic biases in context of PASSION.

- With the advent of telemedicine, developing countries are learning newer ways of leveraging their information and communication technologies (ICTs) to play an increasingly vital role in the health care industry. Telemedicine is defined as a health care delivery mechanism where physicians and other medical personnel can examine patients remotely using information and telecommunication technologies (ICTs; Bashshur, Sanders, and Shannon, 1997). (**Kifle_2024**)
- AI systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations (Mehrabi et al., 2021).
- There are clear benefits to algorithmic decision-making; unlike people, machines do not become tired or bored [45, 119], and can take into account orders of magnitude more factors than people can. However, like people, algorithms are vulnerable to biases that render their decisions “unfair” [6, 121]. In the context of decision-making, fairness is *the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics*. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people. (Mehrabi et al., 2021).
- it is important for researchers and engineers to be concerned about the downstream applications and their potential harmful effects when modeling an algorithm or a system (Mehrabi et al., 2021).
- We should think responsibly, and recognize that the application of these tools, and their subsequent decisions affect peoples’ lives; therefore, considering fairness constraints is a crucial task while designing and engineering these types of sensitive tools (Mehrabi et al., 2021).

2 State of Research

Bezogen auf die eigenen Zielsetzungen und Fragestellungen soll aufgezeigt werden, wie andere dieses oder ähnliche Probleme gelöst haben. Worauf können Sie aufbauen, was müssen Sie neu angehen? Wodurch unterscheidet sich Ihre Lösung von anderen Lösungen? Für wissenschaftlich orientierte Arbeiten sei hier explizit auf (Balzert, S. 66 ff) verwiesen. Relevante, aktuelle und fundierte Fachliteratur wurde identifiziert, kritisch geprüft und verwendet. Die Begriffe der Fragestellung sind definiert und referenziert. Der gesamte Kontext ist verknüpft und eine Abgrenzung wurde vorgenommen. All dies ist in einer leicht verständlichen Struktur formuliert und überprüft.

This chapter provides a review of existing work in the field of bias mitigation in AI. The main focus lies on a literature review of existing papers from other researchers in this area, highlighting the key findings which are connected to this thesis. Bias mitigation in AI has already been investigated by different researchers, who crafted fitting mitigation methods TODO: citation?. This thesis aims to assess those existing methods in the context of PASSION.

Therefore, this chapter first presents an overview of the PASSION project based on the PASSION paper and dataset. Then, the general knowledge in the literature about existing biases, fairness metrics and mitigation methods is summarized. The review process was divided into two main contexts: ML in general, and ML in dermatology. This approach ensures that the technical and dermatological perspectives are considered when applying the knowledge to PASSION. The tables in this chapter indicate which points were found in which context. This is important, since what may be an issue in general might not be relevant for a specific use case or vice versa. For example, in theory, all age groups should be represented in datasets to account for demographic diversity. However, for car insurance, age representation is not important, because age does not affect how well a driver can drive TODO: either cite this example from the expert or find another example related to dermatology.

The various studies present different bias sources and suggest diverse methods to mitigate them. During the literature review, several biases and mitigation methods were identified that may be relevant to the PASSION project. Since it is not feasible to assess all of them during the duration of this thesis, the thesis focuses on those which are related to skin type, age and gender. The chosen methods are explained in chapter 4 Methods. The other items are passed to the PASSION research team as a list for further investigation. The list can be found in the appendix TODO: add link.

TODO: put the evaluation stuff in the execution / analysis section!!

2.1 PASSION for Dermatology

This section provides an overview of the PASSION project regarding its medical scope and technical components.

While the overall goal remains to improve the accessibility of dermatological care by building fair and inclusive AI systems, PASSION specifically addresses common pediatric skin conditions in Sub-Saharan Africa. To create a dataset which represents patients with highly pigmented skin, they collected data from patients with Fitzpatrick skin type (FST) III to VI. Based on this dataset, the PASSION team fine-tuned a ResNet-50 model using transfer learning. With the dataset and trained model, the researchers published data analysis scripts and initial insights on the model performance in a MICCAI TODO: add to glossary publication (Gottfrois et al., 2024).

For the purpose of this thesis, it is essential to understand the dataset's metadata, the architecture and fine-tuning process of the PASSION model and which bias mitigation methods have already been applied. The dataset can influence which biases could arise in the model or rather which ones can be measured. The labels which should be predicted, and the model architecture give insight into the ML task. All this information affects which mitigation methods are feasible to be used for the project. TODO: add sources

2.1.1 PASSION Dataset

The PASSION dataset contains data from patients from four African countries in dermatology clinics. It contains 4901 images of 1653 dermatology cases with the corresponding demographic and clinical metadata. Each patient is represented by one record, with images linked to the record via filename. The images were captured with mobile phones to ensure that the training data complies with a teledermatology setting regarding image quality (Gottfrois et al., 2024).

A predefined 80/20 stratified train-test split at patient level ensures reproducibility and fair comparison, while preventing information leaking (Gottfrois et al., 2024).

Stratified splitting is a method to split datasets while maintaining the original class distribution within the subsets. This is important for imbalanced datasets to maintain minority class representation (Baldé, 2023).

The metadata, as listed in Table 2.1, includes demographic attributes such as *age*, *sex*, and *FST*. These are essential for identifying potential demographic biases lateron. The labels *impetig* and *conditions_PASSION* represent dermatology diagnosis as evaluated by dermatologists (Gottfrois et al., 2024), and are the target variables the PASSION model learns to predict. Therefore, this ML task is a multi-label classification problem. PASSION addresses this by training separate models for each label (Gottfrois et al., 2024). The prediction of conditions_PASSION is a multiclass classification task, while predicting impetig is a binary classification task.

Metadata Attribute	Data Type	Description
subject_id	string	Participant's unique identifier
country	string	Country of data origin
age	integer	Age of the participant in years
sex	m/f/o	Gender of the participant
fitzpatrick	integer	FST
body_loc	string (list; null-able, semicolon-separated)	Specifically affected body locations
impetig	0/1	Presence of impetigo (1=present), may occur alone or with other conditions, affects the treatment options for coexisting conditions
conditions_PASSION	Eczema, Scabies, Fungal, Others	Primary diagnosed skin condition

Table 2.1: PASSION dataset - metadata attributes and descriptions (Gottfrois et al., 2024)

The PASSION team also provides a set of Jupyter Notebook-based data analysis scripts. For example, one script analyses the correlation between the clinical conditions and location of the data collection. A full list of these scripts is included in appendix A PASSION Data Analysis Scripts. Additionally, the paper visualizes demographic analyzes related to age, sex and FST as shown in Figure 2.1.

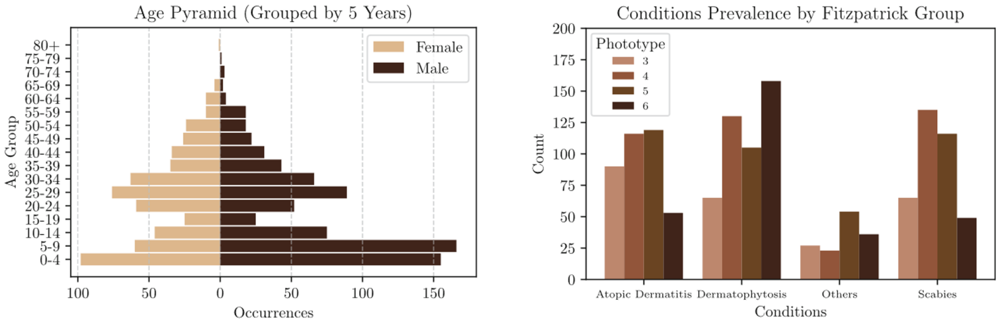


Fig. 1. Age distribution per gender.

Fig. 2. Prevalence per FST.

Figure 2.1: PASSION data distributions (Gottfrois et al., 2024)

Due to the sensitivity of patient data, the dataset is confidential. Access to it can be requested via the project website: https://passionderm.github.io/ (Gottfrois et al., 2024).

2.1.2 PASSION Model

The model architecture is a ResNet-50 model which is pretrained on ImageNet. The model was fine-tuned by replacing the last fully connected classification layer with a dropout layer with a 0.3 dropout rate followed by batch normalization. The class activation is done by a single linear layer. To minimize the weighted cross-entropy loss, Adam optimization is used. For improved generalization and to avoid overfitting, data augmentations were applied. The methods used were random resizing, cropping, flipping, and rotating. For training, the model uses 5-fold cross-validation Gottfrois et al. (2024).

2.1.3 PASSION Experiments

The PASSION team conducted various experiments to evaluate the classifiers on the test set with the following schemes (Gottfrois et al., 2024):

- Performance for skin condition prediction
- Performance for impetigo detection
- Generalization from two centers to a wider population (test set contains data from the known centers and one unknown center)
- Generalization from different age groups (test set contains data from the known age groups and one unknown)
- Subject level analysis over the predictions of multiple pictures, using majority voting

The code for those experiments is available in the PASSION evaluation GitHub repo. This repo can serve as a starting point, since reproducing the results helps to verify that the provided setup works the same on my side. Also, they can be used as examples for further experiments. TODO: mention which ones I really used why for the thesis and move the others to the appendix

The paper indicates lower performance when evaluating the model on a subject level (performance per case/patient) rather than a sample level (performance per image). The authors emphasize the importance of assessing classifier performance on both levels for completeness (Gottfrois et al., 2024). Therefore, the subject level performance should also be considered during this thesis. TODO: challenge this to be tested again in the outlook bc of the inproper metadata linkage

2.1.4 Limitations

TODO: maybe move to execution phase TODO: write in more details - multiple executions showed inconsistent results for the different group evaluations on the same model checkpoint. It turned out that the metadata linkage did not work consistently. I resolved the issue was resolved by providing the image name in the data loader and link the metadata directly from the source file instead of using the indexes. probably related to different shuffling between data loader and metadata loader

2.2 Bias

This chapter provides an overview of biases and related demographic characteristics mentioned in ML- and dermatology-related research. It also explains their relevance for PASSION.

Algorithmic decisions made by AI systems can directly affect peoples' lives. In healthcare applications such as PASSION, these decisions are especially sensitive, as they influence diagnoses and treatment outcomes. Diverse studies have shown that AI application's decisions can hold biases that affect underrepresented groups. This leads to unfair or even harmful consequences. Therefore, it is essential for AI engineers to identify, address, and mitigate such biases in order to develop fair applications. This requires an understanding of what bias is in general, which concrete biases exist, and where they originate (Mehrabi et al., 2021).

2.2.1 Definition of Bias in ML

In the context of ML, bias can be defined as *a systematic error that causes a model or estimator to consistently deviate from the true value or relationship* (Delgado-Rodríguez & Llorca, 2004; Taylor, 2023). In practice, this often results in models that make less accurate predictions for specific subgroups within the population TODO: cite this.

TODO: make sure the following is cited correctly

2.2.2 Demographic Biases in the Context of Dermatology

Biases in dermatology in general can lead to unequal outcomes for different groups, which can result in unfair outcomes for certain groups. Demographic biases are particularly relevant in the context of dermatology AIs, as they can cause differences in diagnostic accuracy and treatment outcomes among different demographic (sub-)groups. From the literature review, three main ways have been identified in which demographic differences may introduce bias in dermatology ML models: TODO: cite all that, from presentation

- **Disease Presentation**. *Skin type* affects how diseases appear on the skin. As Gottfrois notes, "any condition linked to inflammation is less visible if the skin is more pigmented" TODO: cite mail from philippe. This directly influences training and evaluating image-based ML models like those used in PASSION. For example, a model trained predominantly on images with low pigmented skin may perform poorly on images of highly pigmented skin.
- **Disease Prevalence**. Factors such as *age* and *sex* do not tend to affect disease presentation, but they can influence disease prevalence TODO: cite mail from philippe. Also, *geographic location* can influence the prevalence of skin conditions (e.g., tropical vs. dry climates) TODO: add source. Therefore, these factors could introduce bias if certain conditions are underrepresented in the dataset due to demographic imbalances. TODO: consider adding

smt like the car driver example here, indicating that it is not necessarily a problem due to the same disease presentation

- **Access to Healthcare**. *Socioeconomic status* or *geographic location* can also introduce bias. Research shows that patients with lower socioeconomic status are often diagnosed at later stages of the disease, which may alter the visual presentation of the disease. If such cases are missing in training data, the model may fail to recognize them, leading to misdiagnosis. TODO: add example for geographic location?.

To build a robust and fair ML model, it is essential to identify and address biases linked to such protected characteristics (**Mehrabi2022**). TODO: check that there is no duplication between PASSION dataset feature description and here TODO: probably remove Due to time constraints, this thesis focuses on three protected characteristics: skin type, age, and sex. These were selected based on their presence in the PASSION dataset and their influence on dermatological diagnosis and disease prevalence. Other potentially relevant features, such as geographic location and socioeconomic status, should be evaluated in future work by the PASSION team.

2.2.3 Types of Biases and Their Relevance for PASSION

The literature describes numerous types of bias. Over 60 were identified during this research. These factors were grouped into categories to provide an overview, and their relevance to the PASSION context was assessed.

Among them, *sampling biases* and *representation biases* are particularly relevant, as they relate directly to the inclusion or exclusion of demographic subgroups in the dataset. For example, *ascertainment bias*, a subtype of sampling bias, occurs when parts of the target population are unintentionally excluded. A common example is healthcare studies conducted in public hospitals only, which excludes patients from higher socioeconomic backgrounds who visit private clinics. This skews the data and can lead to incorrect conclusions, such as overestimating disease prevalence in specific groups.

Other relevant categories include *medical biases* and *imaging biases*, especially in the teledermatology setting of PASSION. These include clinical labeling errors, variations in image quality or lighting conditions which lead to bias.

This thesis focuses on the most relevant bias types. An extensive list is provided in appendix B List of Biases and will be shared with the PASSION team for further evaluation.

TODO: add the 5-10 most important biases here

2.2.4 Sensitive Features

Research has identified sensitive features that are particularly prone to bias. These features have already caused biases in existing AI applications and should therefore be carefully evaluated during model development (Mehrabi et al., 2021).

Table 2.2 summarizes sensitive features mentioned in the literature. The categorization in the table was done based on the research described in subsection 2.2.2. For completeness, the table also contains sensitive demographic features which appear unrelated to dermatology according based on current research.

Bias-Sensitive Features	**Mentioned in Context of**	
	ML	**Dermatology**
Related to Disease Presentation		
Skin Type	X[1,2,7]	X[12,13]
Skin Undertones		X[13]
Socio-Economic Status	X[6]	X[12]
Geographic Location TODO: double check this!	X[1,3]	
Related to Disease Prevalence		
Age	X[7,11]	X[13]
Gender/Sex	X[1,2,7,8,9,10,11]	X[13]
Gender and Skin Type Subgroups	X[1,2]	
Related to Access to Healthcare		
Geographic Location	X[1,3]	
Socio-Economic Status	X[6]	X[12]
Relation to Dermatology to be Checked		
Ethnicity/Race	X[1,2,4,5,6,7,11]	X[12,13]
Disabilities	X[7,11]	
Unrelated to Dermatology		
Familial status	X[7]	
Marital status	X[7,11]	
Nationality/National origin	X[7,11]	
Recipient of public assistance	X[7]	
Religion	X[7,11]	

[1] (Mehrabi et al., 2021)
[2] (Buolamwini & Gebru, 2018)
[3] (Shankar et al., 2017)
[4] (Manrai et al., 2016)
[5] (Fry et al., 2017)
[6] (Vickers & Fouad, 2014)
[7] (J. Chen et al., 2019)
[8] (Zhao et al., 2017)
[9] (Bolukbasi et al., 2016)
[10] (Zhao et al., 2018)
[11] (Hajian & Domingo-Ferrer, 2013)
[12] (Young et al., 2020)
[13] (Montoya et al., 2025)

Table 2.2: Commonly used features which often are affected by biases

2.3 Fairness Metrics

This chapter introduces the concept of fairness in ML, as fairness is a way to detect whether and what biases exist in a model. As there is no universally accepted definition of fairness, various fairness metrics have been proposed in the literature, each based on different assumptions and goals (Mehrabi et al., 2021).

2.3.1 Definition of Fairness in ML

In research, there is currently no common agreement regarding a fairness definition in ML. Broadly, fairness *is the absence of bias towards individuals or groups in a decision-making context.* To assess how fair AI models are, multiple fairness metrics have been proposed in the literature, each reflecting different interpretations of fairness. The choice of metric largely depends on the specific use case of the application (Mehrabi et al., 2021).

2.3.2 Fairness Metrics

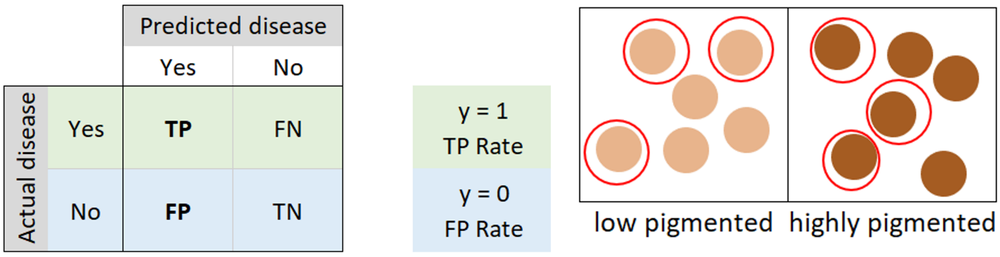
Mehrabi et al. (2021) summarized the fairness metrics and grouped them into the categories group fairness, subgroup fairness and individual fairness, depending on the main mechanics of the metrics. They are listed in Table 2.3.

Fairness Definitions	**Mentioned in Context of**	
	ML	**Dermatology**
Group Fairness		
Conditional Statistical Parity	X	
Demographic/Statistical Parity	X	
Equal Opportunity	X	
Treatment Equality	X	
Test Fairness	X	
Equalized Odds	X	
Subgroup Fairness		
Subgroup Fairness	X	
Individual Fairness		
Counterfactual Fairness	X	
Fairness Through Awareness	X	
Fairness Through Unawareness	X	
Not Categorized		
Fairness in Relational Domains	X	

Table 2.3: Fairness definitions based on Mehrabi et al. (2021)

To better understand how fairness can be formally defined, consider the example of equalized odds, introduced by Hardt et al. (2016):
”*A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y.*
$P(\hat{Y} = 1 \mid A = 0, Y = y) = P(\hat{Y} = 1 \mid A = 1, Y = y), \quad \forall y \in \{0, 1\}$” TODO: add formula list
In other words, the probability of predicting a positive outcome should be the same across protected and unprotected groups, given the true label Y. This ensures that both true positive rate (TPR) and false positive rate (FPR) are equal across different demographic groups. If these rates are the same, like in the example of Figure 2.2, the model satisfies equalized odds, and fairness is achieved.

Since equalized odds compares conditional probability distributions across groups, it is a group fairness metrics.



		Predicted disease	
		Yes	No
Actual disease	Yes	**TP**	FN
	No	**FP**	TN

Figure 2.2: Equalized odds mechanics, inspired by Kearns et al. (2019).

The mechanics of the other fairness metrics are described broadly in appendix C Fairness Metrics.

There are Python libraries like *Fairlearn* available, which can be used for the computation of the fairness metric (Agarwal et al., 2018). They tend to support the most popular metrics for binary classification (contributors, n.d.).

2.3.3 Limitations of Group Fairness

Despite its usefulness, equalized odds and similar group fairness metrics have limitations. These metrics can hide inequalities that exist within more specific subgroups. For example, a model might appear fair when assessed across broad groups such as age or skin type (Figure 2.2) but still exhibit substantial disparities within subgroups, such as older individuals with darker skin tones (Figure 2.3) (Kearns et al., 2018, 2019).

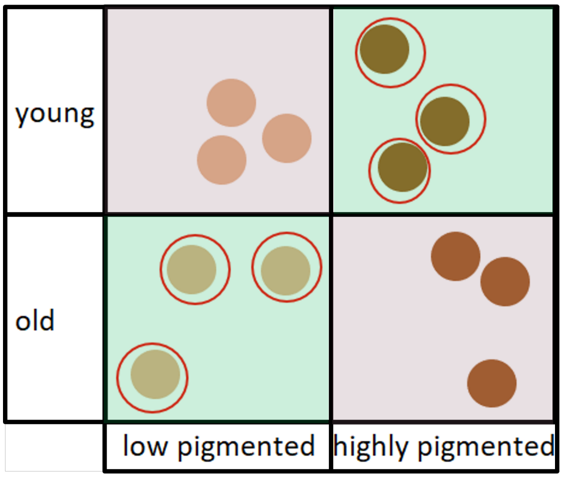


Figure 2.3: Equalized odds violations on subgroups, inspired by Kearns et al. (2019).

To address this issue, subgroup fairness metrics have been proposed. These extend group fairness metrics by explicitly evaluating fairness across subgroups. This ensures that fairness assessments do not overlook hidden biases that could affect smaller populations (Kearns et al., 2018, 2019).

2.4 Mitigation Methods

TODO: still to be written

see text from bias chapter - Further, AI engineers need to know what prevention methods are available to reduce the biases (Mehrabi et al., 2021).

2.4.1 Mitigation Methods Overview

TODO: write definitions of pre-in and post-processing, see Methods for fair machine learning below [43, 11, 14]

TODO: add stratified split TODO: double check and futher improve groups

Mitigation Methods - Unbiasing Data (Pre-Processing)	**Mentioned in Context of**	
	ML	**Dermatology**
Documentation and Transparency		
Good Practices while using Data	X[1,2,3]	
Datasheets as supporting document for dataset creation method, characteristics, motivations and skews	X[1,2,3]	
Datasheets as supporting document for model method, characteristics, motivations and skews	X[1,4]	
Dataset (Nutrition) Labels	X[1,5,6]	X[18, TODO: add spec source]
Communication and Reporting		
Messaging	X[1,12]	
Bias Detection and Evaluation		
Test for Simpson's Paradox TODO: Discribe Simpson's Paradox	X[1,7,8,9]	
Detect Direct Discrimination with Causal Models and Graphs	X[1,10]	
Out-of-Distribution Detection in Dermatology Using Input Perturbation and Subset Scanning		X[19]
Check confidence interval and p-curve analysis instead of p-value		X[17]
Study Design		
Allocation concealment and blinding		X[17]
Preventing Direct and Indirect Discrimination	X[1,11]	
Data Gathering		
Data Collection from diverse sources (incl. primary care clinics)	X[18]	
Robust standards for external validation	X[18]	
Preferential Sampling	X[1,13,14]	
Geographical Diversity and Inclusion for Dataset creation	X[16]	
Balanced Representation accross skin tones and genders		X[19]
Disparate Impact Removal	X[1,15]	
Labeling		
Multidimensional Scale for Skin Tones		X[19]
Data Availability and Open Science		
Publish Datasets accessible for the public		X[18, TODO: add source]

[1] (Mehrabi et al., 2021)
[2] (**M13__**)
[3] (**M55__**)
[4] (**M110__**)
[5] (**M66__**)
[6] (**M66Successor__**)
[7] (**M81__**)
[8] (**M3__**)
[9] (**M4__**)
[10] (**M163__**)
[11] (Hajian & Domingo-Ferrer, 2013)
[12] (**M74__**)
[13] (**M75__**)
[14] (**M76__**)
[15] (**M51__**)
[16] (Shankar et al., 2017)
[17] (Chakraborty, 2024)
[18] (Young et al., 2020)
[19] (Montoya et al., 2025)

Table 2.4: Mitigation Methods - Unbiasing Data - Mentioned in Contextual Research, grouped like in Mehrabi et al. (2021), the author cannot guarantee for completeness

Mitigation Methods - Fair Classification	**Mentioned in Context of**	
	ML	**Dermatology**
Satisfy Fairness Definitions		
Satisfy Subgroup Fairness TODO: unclear if * in [3] as well, or if [2] also handles *	X[1,2]	
Satisfy Equality of Opportunity*	X[1,3,6]	
Satisfy Equalized Odds*	X[1,3]	
Disparate Treatment**	X[1,4,5]	
Disparate Impact**	X[1,4,5]	
TODO: find out exact method	X[1,7]	
TODO: find out exact method	X[1,8]	
TODO: find out exact method	X[1,9]	
TODO: find out exact method	X[1,10]	
Satisfy Fairness and Stability Under Distribution Shifts		
TODO: find out exact method	X[1,11]	
Fair Representation Learning (Pre/In-processing)		
Representation Learning by Disentanglement	X[1,2]	
Variational Fair Autoencoder	X[1,3,15]	
VAE without adversarial training	X[1,4]	
Adversial Learning with FairGAN	X[1,16]	
Removing correlation between protected and unprotected features with a geometric solution	X[1,17]	
Algorithmic Adaptions for Fairness		
Modified Discrimination-Free Naive Bayes Classifier	X[1,12]	
Fairness-Aware ML Frameworks		
Fairness-Aware Classification Framework	X[1,13]	
Fairness Constraints in Multitask Learning (MTL) Framework	X[1,14]	
Decoupled Classification System with Transfer Learning	X[1,15]	
Preferential Data Selection and Representation		
Wasserstein Distance Measure for Dependence Mitigation	X[1,16]	
Preferential Sampling (PS) for Discrimination-Free Training Data	X[1,17]	
Model Interpretability		
Post-Processing with Attention Mechanism	X[1,18]	
Use Brier Score and Response Rate Accuracy		X[19], TODO: add clear source
some more methods TODO: describe		X[19]

* possible to satisfy together

** possible to satisfy together

[1] (Mehrabi et al., 2021)

[2] (**M147__**)

[3] (Hardt et al., 2016)

[4] (**M2__**)

[5] (**M159__**)

[6] (**M154__**)

[7] (**M57__**)

[8] (**M78__**)

[9] (**M85__**)

[10] (**M106__**)

[11] (**M69__**)

[12] (**M25__**)

[13] (**M155__**)

[14] (**M12__**)

[15] (**M49__**)

[16] (**M73__**)

[17] (**M75__**)

[18] (**M102__**)

[19] (Young et al., 2020)

Table 2.5: Mitigation Methods - Fair Classification - Mentioned in Contextual Research, grouped like in Mehrabi et al. (2021), the author cannot guarantee for

TODO: check categorization

Mitigation Methods - not so relevant for us	**Mentioned in Context of**	
	ML	**Dermatology**
Fair NLP		
Fair Word-Embedding	X[1,5,6,7]	
Train-Time Data Augmentation	X[1,8]	
Test-Time Neutralization	X[1,8]	
Fair Regression (In-processing)		
Price of Fairness (POF)	X[1,10]	
XY TODO: check this and bounded group loss	X[1,11]	
Decision Tree for Disparate Impact and Treatment	X[1,12]	
Structured Prediction (In-processing)		
Reducing Bias Amplification (RBA) as calibration algorithm	X[1,13]	
Principal Component Analysis (PCA) (In-processing)		
Fair PCA	X[1,14]	
Graph-Based Fairness Methods		
Community Detection / Graph Embedding TODO: how to proceed with this	X	
Causal Fairness and Disparate Learning		
Disparate Learning Processes (DLP)	X[1,9]	
Causal Approach to Fairness TODO: how to proceed with this	X[TODO: add clear source]	
Disregard path in causal graph which result in sensitive attributes affecting decision outcome	X[1]	
Removing Sensitive Attributes		
Disregard sensitive attributes in effect on decision-making	X[1]	

[1] (Mehrabi et al., 2021)
[2] (**M42_**)
[3] (**M97_**)
[4] (**M112_**)
[5] (Bolukbasi et al., 2016)
[6] (**M58_**)
[7] (**M169_**)
[8] (**M166_**)
[9] (**M94_**)
[10] (**M14_**)
[11] (**M1_**)
[12] (**M2_**)
[13] (Zhao et al., 2017)
[14] (**M137_**)
[15] (**M5_**)
[16] (**M90_**)
[17] (**M65_**)

Table 2.6: Mitigation Methods - Others - Mentioned in Contextual Research, grouped like in Mehrabi et al. (2021), the author cannot guarantee for completeness

TODO: mention also the IBM AI Fairness 360 toolkit [11] and that authors evaluated their work in benchmark datasets [65], [72], [158], [159]

TODO: draft for presentation satisfy Equalized Odds / Subgroup fairness highlight allocation concealment and blinding and data collection from diverse sources and Preferential Sampling

2.4.2 Mitigation Methods Overview

Mitigation Methods	**Mentioned in Context of**	
	ML	**Dermatology**
Unbiasing Data		
Documentation and Transparency	X[1]	X[3]
Bias Detection and Evaluation	X[1]	X[2,4]
Study Design	X[1]	X[2]
Data Gathering	X[1]	X[3,4]
Data Availability and Open Science		X[3]
Removing Sensitive Attributes	X[1]	
Fair Classification		
Satisfy Fairness Definitions	X[1]	
Satisfy Fairness and Stability Under Distribution Shifts	X[1]	
Fair Representation Learning	X[1]	
Fairness-Aware ML Frameworks	X[1]	
Preferential Data Selection and Representation	X[1]	
Model Interpretability	X[1]	X[3]
For Other ML Algorithm Types		
Fair NLP	X[1]	
Fair Regression	X[1]	
Structured Prediction	X[1]	
Fair Principal Component Analysis	X[1]	
Graph-Based Fairness Methods	X[1]	
Causal Fairness and Disparate Learning	X[1]	

[1] (Mehrabi et al., 2021)
[2] (Chakraborty, 2024)
[3] (Young et al., 2020)
[4] (Montoya et al., 2025)

Table 2.7: Mitigation Methods - Draft

TODO: check categorization

Mitigation Methods - not so relevant for us	**Mentioned in Context of**	
	ML	**Dermatology**
Fair NLP		
Fair Word-Embedding	X[1,5,6,7]	
Train-Time Data Augmentation	X[1,8]	
Test-Time Neutralization	X[1,8]	
Fair Regression (In-processing)		
Price of Fairness (POF)	X[1,10]	
XY TODO: check this and bounded group loss	X[1,11]	
Decision Tree for Disparate Impact and Treatment	X[1,12]	
Structured Prediction (In-processing)		
Reducing Bias Amplification (RBA) as calibration algorithm	X[1,13]	
Principal Component Analysis (PCA) (In-processing)		
Fair PCA	X[1,14]	
Graph-Based Fairness Methods		
Community Detection / Graph Embedding TODO: how to proceed with this	X	
Causal Fairness and Disparate Learning		
Disparate Learning Processes (DLP)	X[1,9]	
Causal Approach to Fairness TODO: how to proceed with this	X[TODO: add clear source]	
Disregard path in causal graph which result in sensitive attributes affecting decision outcome	X[1]	
Removing Sensitive Attributes		
Disregard sensitive attributes in effect on decision-making	X[1]	

[1] (Mehrabi et al., 2021)
[2] (**M42__**)
[3] (**M97__**)
[4] (**M112__**)
[5] (Bolukbasi et al., 2016)
[6] (**M58__**)
[7] (**M169__**)
[8] (**M166__**)
[9] (**M94__**)
[10] (**M14__**)
[11] (**M1__**)
[12] (**M2__**)
[13] (Zhao et al., 2017)
[14] (**M137__**)
[15] (**M5__**)
[16] (**M90__**)
[17] (**M65__**)

Table 2.8: Mitigation Methods - Others - Mentioned in Contextual Research, grouped like in Mehrabi et al. (2021), the author cannot guarantee for completeness

2.4.3 Mitigation Methods Extensive Sources

Bias Examples and Mitigation Ideas

Data bias examples and mitigation ideas

- Bias in ML Data - (Buolamwini & Gebru, 2018) IJB-A / Adience imbalanced (mainly light-skinned subjects) - Bias towards dark-skinned groups (underrepresented). Other instance - when we do not consider different subgroups in the data. Considering only male-female groups not enough, use race to further subdivide gender groups. Only then, clear biases in sub groups can be found, since otherwise part of the groups would compromise the other group and hide the underlaying bias towards that subgroup (Mehrabi et al., 2021)
- Popular machine-learning datasets that serve as a base for most of the developed algorithms and tools can also be biased—which can be harmful to the downstream applications that are based on these datasets. ... In [(Shankar et al., 2017), researchers showed that these datasets suffer from representation bias and advocate for the need to incorporate geographic diversity and inclusion while creating such datasets. (Mehrabi et al., 2021)
- Examples of Data Bias in Medical Applications. These data biases can be more dangerous in other sensitive applications. For example, in medical domains there are many instances in which the data studied and used are skewed toward certain populations—which can have dangerous consequences for the underrepresented communities. [98] showed how exclusion of African-Americans resulted in their misclassification in clinical studies, so they became advocates for sequencing the genomes of diverse populations in the data to prevent harm to underrepresented populations (Mehrabi et al., 2021) TODO: What does sequencing data mean?, is it relevant

Methods for Fair Machine Learning

- While this section is largely domain-specific, it can be useful to take a cross-domain view. Generally, methods that target biases in the algorithms fall under three categories (Mehrabi et al., 2021)
- Pre-processing. Pre-processing techniques try to transform the data so that the underlying discrimination is removed [43]. If the algorithm is allowed to modify the training data, then pre-processing can be used [11].(Mehrabi et al., 2021)
- In-processing. In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process [43]. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used during the training of a model— either by incorporating changes into the objective function or imposing a constraint [11, 14].(Mehrabi et al., 2021)
- Post-processing. Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model [43]. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model

initially get reassigned based on a function during the post-processing phase [11, 14].(Mehrabi et al., 2021)

- we concentrate on discrimination prevention based on preprocessing, because the preprocessing approach seems the most flexible one: it does not require changing the standard data mining algorithms, unlike the inprocessing approach, and it allows data publishing (rather than just knowledge publishing), unlike the postprocessing approach. (Hajian & Domingo-Ferrer, 2013) –> TODO: this is an important point which we should consider for PASSION, also, some more insight in regards of the different phases can be found in this paper
- From learning fair representations [42, 97, 112] to learning fair word embeddings [(Bolukbasi et al., 2016), 58, 169], debiasing methods have been proposed in different AI applications and domains. (Mehrabi et al., 2021) –> seems to refer mostly to NLP domains
- Most of these methods try to avoid unethical interference of sensitive or protected attributes into the decision-making process, while others target exclusion bias by trying to include users from sensitive groups. (Mehrabi et al., 2021)
- However, a recent paper [58] argues against these debiasing techniques and states that many recent works on debiasing word embeddings have been superficial, that those techniques just hide the bias and don't actually remove it. (Mehrabi et al., 2021)
- some works try to satisfy one or more of the fairness notions in their methods, such as disparate learning processes (DLPs) which try to satisfy notions of treatment disparity and impact disparity by allowing the protected attributes during the training phase but avoiding them during prediction time [94].(Mehrabi et al., 2021)
- Some of the existing work tries to treat sensitive attributes as noise to disregard their effect on decision-making, while some causal methods use causal graphs, and disregard some paths in the causal graph that result in sensitive attributes affecting the outcome of the decision.(Mehrabi et al., 2021)
- Different bias-mitigating methods and techniques are discussed below for different domains—each targeting a different problem in different areas of machine learning in detail. (Mehrabi et al., 2021)

Unbiasing Data

- Every dataset is the result of several design decisions made by the data curator. Those decisions have consequences for the fairness of the resulting dataset, which in turn affects the resulting algorithms. In order to mitigate the effects of bias in data, some general methods have been proposed that advocate having good practices while using data, such as having datasheets that would act like a supporting document for the data reporting the dataset creation method, its characteristics, motivations, and its skews [13, 55]. A similar suggestion has been proposed for models in [110].(Mehrabi et al., 2021)

- Authors in [66] also propose having labels, just like nutrition labels on food, in order to better categorize each data for each task. (Mehrabi et al., 2021)
- some work has targeted more specific types of biases. For example, [81] has proposed methods to test for cases of Simpson's paradox in the data, and [3, 4] proposed methods to discover Simpson's paradoxes in data automatically. (Mehrabi et al., 2021)
- Causal models and graphs were also used in some work to detect direct discrimination in the data along with its prevention technique that modifies the data such that the predictions would be absent from direct discrimination [163].(Mehrabi et al., 2021)
- in [(Hajian & Domingo-Ferrer, 2013)] also worked on preventing discrimination in data mining, targeting direct, indirect, and simultaneous effects.(Mehrabi et al., 2021)
- Other pre-processing approaches, such as messaging [74], preferential sampling [75, 76], disparate impact removal [51], also aim to remove biases from the data. (Mehrabi et al., 2021)
- Image quality. Several barriers to AI implementation in the clinic need to be overcome with regards to imaging (Figure 1). These include technical variations (e.g., camera hardware and software) and differences in image acquisition and quality (e.g., zoom level, focus, lighting, and presence of hair). For example, the presence of surgical ink markings is associated with decreased specificity (Winkler et al., 2019), field of view can significantly affect prediction quality (Mishra et al., 2019), and classification performance improves when hair and rulers are removed (Bisla et al., 2019). We have developed a method to measure how model predictions might be biased by the presence of a visual artifact (e.g., ink) and proposed methods to reduce such biases (Pfau et al., 2019). Poor quality images are often excluded from studies, but the problem of what makes an image adequate is not well studied. Ideally, models need to be able to express a level of confidence in a prediction as a function of image quality and appropriately direct a user to retake photos if needed. (Young et al., 2020) - dermatology

Fair Classification

- certain methods have been proposed [57, 78, 85, 106] that satisfy certain definitions of fairness in classification. For instance, in [147] authors try to satisfy subgroup fairness in classification, equality of opportunity and equalized odds in [63], both disparate treatment and disparate impact in [2, 159], and equalized odds in [154]. (Mehrabi et al., 2021)
- Other methods try to not only satisfy some fairness constraints but to also be stable toward change in the test set [69] (Mehrabi et al., 2021)
- The authors in [155], propose a general framework for learning fair classifiers. This framework can be used for formulating fairness-aware classification with fairness guarantees. In another work [25], authors propose three different

modifications to the existing Naive Bayes classifier for discrimination-free classification.(Mehrabi et al., 2021)

- paper [122] takes a new approach into fair classification by imposing fairness constraints into a Multitask learning (MTL) framework. In addition to imposing fairness during training, this approach can benefit the minority groups by focusing on maximizing the average accuracy of each group as opposed to maximizing the accuracy as a whole without attention to accuracy across different groups. In a similar work [49], authors propose a decoupled classification system where a separate classifier is learned for each group. They use transfer learning to reduce the issue of having less data for minority groups.(Mehrabi et al., 2021)
- In [73] authors propose to achieve fair classification by mitigating the dependence of the classification outcome on the sensitive attributes by utilizing the Wasserstein distance measure.(Mehrabi et al., 2021)
- In [75] authors propose the Preferential Sampling (PS) method to create a discrimination free train data set. They then learn a classifier on this discrimination free dataset to have a classifier with no discrimination.(Mehrabi et al., 2021)
- In [102], authors propose a post-processing bias mitigation strategy that utilizes attention mechanism for classification and that can provide interpretability. (Mehrabi et al., 2021)

Fair Regression TODO: only summarize briefly, as PASSION is a classification and not a regression task

- “price of fairness” (POF) to measure accuracy-fairness trade-offs, 3 penalites: Individual fairness, group fairness and hybrid fairness [14] (Mehrabi et al., 2021)
- In addition to the previous work, [1] considers the fair regression problem formulation with regards to two notions of fairness statistical (demographic) parity and bounded group loss. [2] uses decision trees to satisfy disparate impact and treatment in regression tasks in addition to classification. (Mehrabi et al., 2021)

Structured Prediction TODO: only summarize briefly, as PASSION is a classification task

- RBA (reducing bias amplification) as calibration algorithm to prevent risk of leveraging social bias, distributions in training data are followed in the predictions. multi-label obeject and visual semantic role labeling classification amplify existing bias in data [(Zhao et al., 2017)] (Mehrabi et al., 2021) –> TODO: be careful with this if the approach would be to generate new images for training!!

Fair PCA TODO: only summarize briefly, as PASSION is a classification task with only like 10 features

- Pincipal Component Analysis (PCA) https://www.geeksforgeeks.org/principal-component-analysis-pca/ –> dimensionality reduction, statistical technic, high-dimensional data into lower-dimensional space while maximising variance in new space -> most important patterns and relationships is preserved
- vanilla PCA exaggerate error in reconstruction for one group of people [137] (Mehrabi et al., 2021)
- And their proposed algorithm is a two-step process listed below: (1) Relax the Fair PCA objective to a semidefinite program (SDP) and solve it. (2) Solve a linear program that would reduce the rank of the solution. [137] (Mehrabi et al., 2021)

Community Detection TODO: use this as an example for out of scope text, - Ludovic approved Community detection algorithms are specifically tailored to analyze network data and find connections in such datasets. For example, they can be used to detect groups of people with similar interest in social networks (Jayawickrama, 2021). This kind of data is not found in the context of PASSION, which is a classification task. Please refer to Mehrabi et al. (2021) for more information on bias mitigation in community detection algorithms.

Causal Approach to Fairness TODO: only relevant, if our variables have a dependency on the variables, e.g. age / gender determines how the disease is presenting itself in the images; check (Mehrabi et al., 2021) page 18 if relevant

Fair Representation Learning https://medium.com/superlinear-eu-blog/representation-learning-breakthroughs-what-is-representation-learning-5dda2e2fed2e

- Variational Auto encoders –> Variational Fair Autoencoder introduced in [97]. Here,they treat the sensitive variable as the nuisance variable, so that by removing the information about this variable they will get a fair representation. They use a maximum mean discrepancy regularizer to obtain invariance in the posterior distribution over latent variables. Adding this maximum mean discrepancy (MMD) penalty into the lower bound of their VAE architecture satisfies their proposed model for having the Variational Fair Autoencoder.
  In [5] authors also propose a debiased VAE architecture called DB-VAE which learns sensitive latent variables that can bias the model (e.g., skin tone, gender, etc.) and propose an algorithm on top of this DB-VAE using these latent variables to debias systems like facial detection systems.
  In [112] authors model their representation-learning task as an optimization objective that would minimize the loss of the mutual information between the encoding and the sensitive variable. The relaxed version of this assumption is shown in Equation 1. They use this in order to learn fair representation and show that adversarial training is unnecessary and in some cases even counter-productive.
  In [42], authors introduce flexibly fair representation learning by disentanglement that disentangles information from multiple sensitive attributes. Their

flexible and fair variational autoencoder is not only flexible with respect to downstream task labels but also flexible with respect to sensitive attributes. They address the demographic parity notion of fairness, which can target multiple sensitive attributes or any subset combination of them. (Mehrabi et al., 2021)

- Adversarial Learning - In [90] authors present a framework to mitigate bias in models learned from data with stereotypical associations. using adversarial networks by introducing FairGAN which generates synthetic data that is free from discrimination and is similar to the real data. They use their newly generated synthetic data from FairGAN, which is now debiased, instead of the real data for training and testing. They do not try to remove discrimination from the dataset, unlike many of the existing approaches, but instead generate new datasets similar to the real one which is debiased and preserves good data utility. (Mehrabi et al., 2021) TODO: address challenges in creating synthetic data in dermatology?

Fair NLP TODO: for PASSION irrelevant, if it wants to stick to ResNet50 Architecture (Gottfrois et al., 2024) and not use Visual Encoders, which would make sense bc of the small dataset

- Word Embedding TODO: potentially relevant, if the labels are used in training, e.g. age / gender determines how the disease is presenting itself in the images; check (Mehrabi et al., 2021) page 21 if relevant
- Coreference Resolution "Coreference resolution involves identifying when two or more expressions in a text refer to the same entity, be it a person, place, or thing." https://medium.com/@datailm/the-key-to-unlocking-true-language-understanding-coreference-resolution-c01d569e2e87 TODO: irrelevant for the PASSION Context

comparison of different mitigation algorithms

- The field of algorithmic fairness is a relatively new area of research and work still needs to be done for its improvement. With that being said, there are already papers that propose fair AI algorithms and bias mitigation techniques and compare different mitigation algorithms using different benchmark datasets in the fairness domain. For instance, authors in [65] propose a geometric solution to learn fair representations that removes correlation between protected and unprotected features. The proposed approach can control the trade-off between fairness and accuracy via an adjustable parameter. In this work, authors evaluate the performance of their approach on different benchmark datasets, such as COMPAS, Adult and German, and compare them against various different approaches for fair learning algorithms considering fairness and accuracy measures [65, 72, 158, 159]. In addition, IBM's AI Fairness 360 (AIF360) toolkit [11] has implemented many of the current fair learning algorithms and has demonstrated some of the results as demos which can be utilized by interested users to compare different methods with regards to different fairness measures. (Mehrabi et al., 2021)

2.4.4 Statistical biases

https://data36.com/statistical-bias-types-explained/

-

2.4.5 Dermatology Bias

- https://ijdvl.com/biases-in-dermatology-a-primer/ 29 biases, 4 reasons to know about it, 7 mitigation methods (Chakraborty, 2024) - dermatology
- A recent study reported mean top-1 and top-5 model accuracy of 44.8% and 78.1%, respectively, for the classification of 134 diseases (Han et al., 2019b). Most datasets are proprietary, often with minimal description, and datasets collected in dermatology clinics may be skewed toward more complex cases, to those patients with better access to care, or by the choice of camera used in one clinic versus another. Data should be collected from as many diverse sources as possible, including primary care clinics, and robust standards for external validation are needed. (Young et al., 2020)
- There have been successful efforts to support reproducibility and open access. For example, the study by Han et al. (2018a) details the number and characteristics of images from each data source and makes thumbnails of the images publicly available. Additionally, several studies classifying dermoscopic images use the publicly available International Skin Imaging Collaboration archive (Gutman et al., 2016). By making datasets public, it becomes possible to examine them for bias (Bissoto et al., 2019). Alternatively, reporting a model training database's patient demographics and disease classes would be helpful in predicting model performance on external populations. (Young et al., 2020)
- Metrics of model performance. Standard metrics are needed to assess the performance of different models (Figure 1). Currently, standard performance metrics such as accuracy and area under the receiver operating characteristic and precision recall curves are routinely reported. However, for use in the clinic, studies should additionally describe how well their models deal with uncertainty by reporting (i) the Brier Score, or mean-squared calibration error (Rufibach, 2010), which measures how reliably a model can forecast its accuracy, and (ii) area under the response rate accuracy curve, which measures how capably a model can identify examples it is likely to predict falsely and thus abstain from predicting (Hendrycks et al., 2019) (Young et al., 2020)
- Model interpretability. Acceptance of AI in clinical decision-making hinges on being able to understand the decisionmaking process fundamental to its predictions. DL models are inherently difficult to interpret because they are complex, routinely containing millions of learned parameters; interpretation of DL models' output is an active field of research (Murdoch et al.,

2019). One approach for interpreting model diagnoses is contentbased image retrieval, a method for retrieving training images that are visually similar to a test image (Tschandl et al., 2019a). This method may reassure the physician if all the retrieved training images have the same diagnosis as the predicted diagnosis but is less helpful if the test image looks similar to two or more training images with conflicting diagnoses. A second approach is to highlight pixels in an image most relevant for a model's prediction, using methods such as saliency mapping (Figure 1). However, it is often the case that highlighted pixels correspond to the entire lesion or visually distinctive features that are already obvious to clinicians without indication as to why these pixels are important to the diagnosis. A third approach is to see through the eyes of a model by plotting an activation atlas (Carter et al., 2019), which shows how subtle changes, in particular visual features, may tip the model over into choosing one diagnosis over another. These activation atlases are experimental and have yet to be applied in dermatology. Understanding a model's predictions and how the prediction is applicable to the patient at hand is necessary to build trust. As AI exceeds human performance in various tasks, interpreting models may help to advance scientific knowledge by understanding what the machine sees that is relevant to its predications (Young et al., 2020)

2.4.5.1 Demographic Bias in Dermatology

fairness melanoma detection

- Some biases can be easily detected and countered, such as through appropriate data curation; for example, having a balanced representation across skin tones and genders in training sets. However, in other cases, biases are hidden and untraceable [9]. (Montoya et al., 2025)
- whether information on demographic diversity (age, gender, race, or ethnicity of patients), clinical diversity (skin type, lesion type, anatomical location of lesion), or image characteristics (source, imaging techniques, resolution, and whether the images were real or artificially generated) (Montoya et al., 2025)
- The most popular skin color scale currently being used for data annotation for image recognition techniques is the Fitzpatrick Skin Tone Scale (FST) [10]which has six skin tones. Dating from the 1970s, it originally featured just 4 light tones and was designed for detecting photo sensitivity for white skin, with two darker tones added later [11]. The Monk Skin Scale was recently developed and still needs testing, but promisingly has 10 tones, 5 light and 5 dark [12].(Montoya et al., 2025) TODO: highlight this (FST alternatives)
- Fig. 4. Comparison of skin tone scales that can be used for skin cancer detection utilizing AI. Recreation of fitzpatrick skin type scale, monk skin tone scale, and sampling of L'Oreal color chart map for reference. (Montoya et al., 2025) TODO: include this figure
- While this systemic review provides a comprehensive review of the literature

on fairness in AI for melanoma detection, it is primarily based on existing research. To validate the proposed recommendations or frameworks, continuing work is necessary to complete empirical analysis and experiments. Additionally, the suggested adoption of new skin tone scales, while beneficial, may face practical challenges in implementation. Furthermore, while the paper strongly advocates for specific skin tone scales, it's important to note that other methods or tools might also effectively address fairness issues in AI for melanoma detection. Finally, while the study addresses fairness in AI, it could benefit from further exploration of the practical implementation of these recommendations in real-world clinical settings. Potential obstacles and the feasibility of widespread adoption should be considered to ensure that the proposed solutions are not only theoretically sound but also practically viable. (Montoya et al., 2025) TODO: also mention the limitations regarding FST alternatives

- Recent research [13] adds another axis, skin hue, which is described as ranging from red to yellow. This offers a more complete representation of variations of skin color by providing a multidimensional scale [13]. (Montoya et al., 2025)
- The effect of hue (blue, red, yellow, green) on skin tones adds depth to each face producing a range of undertones (cold, neutral, warm, and olive). In the realm of color theory, the concept of 'contrast of hue' emphasizes the distinctiveness among fundamental colors, with primary hues like yellow, red, and blue exhibiting the most pronounced differences [14]. Because skin cancer appears differently on different colored skin, it is important to acknowledge a full range of colors present in both healthy skin and suspicious lesions within datasets used to train skin cancer detection ML tools. (Montoya et al., 2025)
- These findings should correlate to AI for melanoma detection since the contrast between skin color and skin lesions is a preliminary marker during feature extraction. Although the Fitzpatrick Skin Tone (FST) FST measurement scale is not diverse enough and leads to biased AI tools, it is continually used and has even been used to test a recently FDA-approved AI device for detecting melanoma. (Montoya et al., 2025)
- We advocate for the adoption of improved scales like the Monk and L'Oreal maps. Future studies should ensure equitable representation and testing across skin tones to guarantee AI's effectiveness for all. Please refer to Tables 2 through 7 in the discussion section for further recommendations for curating a diverse dataset, including purpose, ownership, funding, and data annotation, as well as recommendations for each stage of the data life cycle. (Montoya et al., 2025) TODO: Link for further mitigation methods
- This study found that while using skin tone instead of race for fairness evaluations in computer vision seems objective, the annotation process remains biased by human annotators. Untested scales, unclear procedures, and a lack of awareness about annotator backgrounds and social context significantly influence skin tone labeling. This study exposes how even minor design choices in the annotation process, like scale order (dark to light in-

stead of light to dark) or image context (face or no face, skin lesion presence), can sway agreement and introduce uncertainty in skin tone assessments. ... The researchers emphasize the need for greater transparency, standardized procedures, and careful consideration of annotator biases to mitigate these challenges and ensure fairer and more robust evaluations in computer vision. (Montoya et al., 2025) - demographic dermatology bias

3 Ideas and Concepts

Hier geht es um die Fragestellung, wie Sie die formulierten Ziele der Arbeit erreichen wollen. Sie halten z.B. erste, grobe Ideen, skizzenhafte Lösungsansätze fest. Gibt es mehrere Wege, Ansätze um dieses Ziel zu erreichen, begründen Sie hier, warum Sie einen bestimmten Weg einschlagen. Beispiel für ein Softwareprojekt: Erste Gedanken über eine grobe Systemarchitektur. Ist z.B. eine Microservice-Architektur angebracht? Welche Alternativen bestehen, wo gibt es Problempunkte? Die Umsetzung, die Beurteilung der Machbarkeit und die detaillierte Beschreibung der umgesetzten Architektur sind dann Teil der Realisierung.

This chapter outlines initial thoughts and conceptual considerations for addressing potential biases in the PASSION project. It sketches the general methodology used in this thesis.

3.1 Broad Methodology

The evaluation and mitigation of bias in the PASSION model is planned to consists of four stages:

1. **Literature Review.** A literature review will be conducted to get an overview of what biases, fairness metrics and mitigation strategies are known in medical AI.
2. **Contextualization and Scope Definition.** The findings' relevance for PASSION's teledermatology context will be evaluated. Based on this, relevant types of bias, applicable fairness metrics and mitigation methods will be selected. Aspects not feasible to address within the scope of this thesis will documented for future work.
3. **Baseline Fairness Assessment.** The current PASSION model will be evaluated using the selected fairness metrics. This will provide a baseline for comparison after mitigation methods are applied.
4. **Mitigation and Evaluation.** Selected mitigation strategies will be implementend individually. Their effect on model fairness and performance will assessed relative to the baseline.

3.2 PASSION Dataset Assessment

In order to decide about the scope and feasibility of the findings in the literature review, the dataset must be assessed. The PASSION dataset was created to improve the representation of highly pigmented skin, which is underrepresented in many traditional dermatology datasets. Nevertheless, it may still lack adequate representation of specific subgroups. Such gaps in representativeness could potentially lead to biased model outputs. However, as Mehrabi et al. (2021) states, this is not necessarily the case. Therefore, a detailed assessed for representativeness can be postponed until the model output indeed proofs to be biased.

Furthermore, the available metadata determines which biases can identified and what mitigation methods are possible. E.g., if metadata on age is missing, fairness with respect to age cannot be assessed.

Therefore, the dataset will be reviewed with regards to:

- Representation of the main groups to get a first impression
- Representation of relevant subgroups if the model output proves to be biased
- Completeness of metadata relevant for fairness evaluation
- Presence of proxy variables that might complicate fairness assessments

These aspects will help determine the extent to which the dataset supports meaningful fairness analysis and subgroup-level model evaluation. It also provides guidance on how to potentially adapt the dataset in the future.

4 Methods

Hier halten Sie fest und begründen, welches Vorgehensmodell Sie für Ihr Projekt wählen. Sie verweisen allenfalls auf die daraus entstandenen, konkreten Terminpläne mit Meilensteinen, welche z.B. unter Realisierung (Kapitel 5) oder im Anhang versorgt sind. Bei Projekten mit einer verlangten wissenschaftlichen Tiefe werden hier die geplanten Forschungsmethoden wie quantitative/qualitative Interviews, Befragungen, Beobachtungen, Feldexperiment etc. beschrieben und begründet. Warum ist in Ihrer Situation ein Interview besser als eine Umfrage? Wer soll interview werden? Die gewählten Methoden sind nachvollziehbar und begründet. Eine methodische Übersicht (Methodisches BigPicture) wurde aufgezeigt und Abgrenzungen erläutert.

This chapter describes the methodological approach and project organization used in this thesis. It outlines the selected process model, planned research methods, and relevant conditions. The focus lies on ensuring that the chosen methods are appropriate, transparent, and justified in the context of evaluating and mitigating bias in the PASSION project.

4.1 Project Management

This chapter illustrates the used process model, how the progress and risk are managed and what technical constraints are available, to get a sense of the constraints and the general plan of this thesis.

4.1.1 Process Model

The project follows the waterfall model. This means the work is done sequentially and each sequence is based on the one before (Petersen et al., 2009). This model has been chosen for the project, since it provides a solid base for the main project while keeping the project management overhead small. This project is separated in two phases:

Phase 1 – Literature Review and Methodology Planning. This phase includes the literature review, the selection and justification of fairness metrics and bias mitigation techniques, and the assessment of the dataset's structure and limitations. Based on these results, a detailed plan for the second phase is developed.

Phase 2 – Execution and Evaluation. In the second phase, the planned assessments and mitigation strategies are implemented. The PASSION model is evaluated against the selected fairness metrics, and improvements are measured and discussed.

The detailed project plan is included in the provided zip-file.

4.1.2 Progress Monitoring and Risk Management

To ensure project transparency and timely delivery, bi-weekly status meetings with the advisor are scheduled. Each meeting is prepared beforehand. Discussed are:

- Work completed in the last period
- Planned work for the next period
- Current project status and comparison with planned schedule
- Top three project risks and planned mitigation strategies

Meeting protocols, including the risk reports are included in the appendix. TODO: add to appendix

4.1.3 Technical Constraints

Model training is performed on HSLU's GPUhub infrastructure, while code development is carried out on a personal notebook. The code is written in Python and builds upon the existing PASSION project architecture. The code base for this thesis is a fork of the PASSION GitHub Project.

- Original Project: https://github.com/Digital-Dermatology/PASSION-Evaluation
- Fork: https://github.com/teshi24/PASSION-Bias-Evaluation

4.2 Literature Review

The literature review targets known bias types, fairness metrics, and mitigation techniques in medical AI, with special attention to teledermatology and demographic factors. Sources include scientific publications, surveys, and technical documentation of relevant libraries. The goal is to build a conceptual and methodological foundation for subsequent analysis.

To ensure the thesis follows scientific standards while still being feasible, the literature review is conducted based on the pragmatic method of Alake (2021) as suggested by my advisor. First, the focus is on survey and taxonomy papers, which provide an overview over the existing research. Them, more detailed papers in the area of dermatology AI is conducted to get more insight in the healthcare context. Such a 2-step approach has also been done by F. Chen et al. (2024). In general, the papers are filtered by focusing on title, abstract and conclusion. Only relevant papers are read in full. TODO: cite protocol in appendix, week1

4.3 Contextualization and Scope Definition

The relevance of the literature findings is evaluated in the context of the PASSION project. This includes analyze the findings from the literature review in terms of their relevance to teledermatology and similar healthcare applications, taking into account the available metadata in the PASSION dataset. Limitations due to dataset constraints or the available time are documented for future work.

The relevance will be categorized into the following groups:

- **High.** Directly applicable to PASSION, both in terms of the teledermatology setting and available metadata; likely to provide valuable insights or improvements.
- **Medium.** Generally relevant to diagnostic AI, but requires adaptations of the PASSION metadata or project in general to be feasible.
- **Low.** Related to PASSION, but only limited.
- **Not Applicable.** Not relevant for PASSION due to fundamental differences in domain, type of data, or type of model.

Based on this contextual analysis, the highly relevant bias types and mitigation methods are investigated further using the most relevant fairness metrics. The selection process follows domain-specific requirements identified in the literature. Such considerations guide the identification of suitable metrics, which are then justified and evaluated in detail during the execution phase.

This contextual analysis is important, as the context and application of fairness metrics and as well as the effect and therefore importance of potential biases can vary by the use case of the AI application (Barr et al., 2025; Mehrabi et al., 2021).

4.4 PASSION Dataset Assessment

The assessment of the PASSION dataset focuses on four core areas:

- **Metadata Completeness.** The metadata is reviewed to verify that all relevant demographic attributes, as identified in the contextualized literature review, are included. Missing attributes limit bias detection and mitigation strategies. They should be added to enable a thorough fairness analysis and bias mitigation. Therefore, potentially missing attributes are listed and passed on to the PASSION team for inclusion in the metadata.
  Further, the available sensitive attributes are identified to ensure that they are included in the subgroup fairness evaluation.
- **Presence of Proxy Variables.** Available metadata attributes are assessed regarding their intended purpose and potential use as proxy variables. If proxy variables are identified, alternatives are proposed to be added to the data instead. This step is essential, as relying on proxy variables may introduce unintended bias into the analysis or model.

- **Representation of Main Groups.** To evaluate overall demographic distributions, the proportions of the values for each demographic attribute (age, sex, FST) are analyzed to identify over- or underrepresented groups. This provides an initial indication of potential data skews, which then can be compared to the model's fairness assessment results. This grants first insight into whether potential unfairness stems from representation bias or other factors.
- **Representation of Relevant Subgroups.** If the fairness assessment of model outputs reveals unfairness on subgroup levels, the distribution of the subgroups is examined using the same method as for the main groups. As this is a more detailed analysis than the representation of main groups, it is done later in the process if biases regarding subgroups in the model indeed exist.

TODO: cite methods

4.5 Reproducing PASSION Results

Before starting any evaluation on the model, the PASSION experiments must be reproduced on the GPUhub, to ensure, that the code base and the data loading is working the same way as for the initial paper. Only then, the evaluation outcome can be used by the PASSION team.

4.6 Fairness Assessment

To establish a reproducible foundation for fairness evaluations within the PASSION project, a baseline fairness assessment with the original project setup is done. For the assessment, the fairness metric selected in subsection 5.1.3 is implemented to analyze model performance across sensitive subgroups, to identify any potential biases in the model output.

The same fairness assessment process is used to evaluate fairness on the model after applying each mitigation method. This ensures consistency and comparability of results across all experimental stages. A mitigation method is considered to hold potential if it significantly improves the fairness assessment results compared to the established baseline.

The fairness is assessed on a subgroup level. For its computation, the Fairlearn library is used where supported to use the standard implementation. Since Fairlearn does not support multiclass analysis and multiple subgroup combinations out of the box, custom code must be developed to handle that part. The subgroups are defined by all unique combinations of the sensitive PASSION metadata attributes as evaluated using the method in chapter 4.4 PASSION Dataset Assessment.

The assessment is run based on the prediction outputs and linked metadata generated in the model evaluation phase, which are cached for later inspection. An independent evaluation class computes the required statistics, and reports fairness metrics. This implementation allows evaluation to be performed independently of model training and supports reproducibility of results.

Alongside with Fairlearn, the implementation builds upon *pandas* and *numpy* for data handling, and *scikit-learn* for standard evaluation metrics.

4.6.1 Limitations

This method provides an initial understanding of fairness in the model output and potential mitigation impacts. However, for scientifically robust conclusions on the fairness impact of a mitigation method, more systematical testing is required.

Ideally, multiple training and evaluation runs per mitigation method using different random seeds should be conducted. Also, the baseline assessment should be run multiple times, using the same seeds to ensure comparison. This approach ensures statistically significant results and accounts for variance due to randomization at diverse stages in the model training process. For instance, Valentim et al. (2019) ran each configuration 30 times using different random seeds.

Due to technical limitations and time constraints, multiple runs were not feasible during this thesis. It is strongly recommended that the PASSION team executes the experiments with additional seeds using the provided scripts, to get a more established result.

4.7 Mitigation Method Evaluation

The PASSION model uses a predefined train-test split. To prevent test set leakage and overfitting while applying mitigation methods, the training data is further divided into a training and a validation set.

If a mitigation method can be applied in multiple ways (e.g., with different parameters, configurations, or data splits), all these variants are evaluated using the train-validation split to prevent test data leakage. The training for all variants will be done without 5-fold cross-validation which allows for significantly faster iteration cycles. This is crucial given the time limitation for this thesis. The variant that performs best on the validation set is then used to evaluate the effectiveness of the mitigation method on the original test set. For this final assessment, 5-fold cross-validation, as setup by Gottfrois et al. (2024), is used again.

This approach ensures that the final test results are comparable across different methods, while keeping the selection process short and independent of the test data.TODO: cite AI lectures

Selected bias mitigation strategies are applied to this setup individually, so that the impact on the fairness can be clearly assigned to the tested mitigation strategy. The impact is evaluated relative to the established baseline as described in chapter 4.6 Fairness Assessment.

To get insight on how the mitigation method influences model performance, also the performance should be compared to the baseline.

4.8 Stratified Split Experiment

Stratified splitting is a bias mitigation method commonly applied using the target labels to ensure a balanced representation of classes. However, additional variables can also be included to maintain minority subgroup representation across train and test sets (Baldé, 2023).

This experiment investigates how different stratification strategies affect model fairness. While the PASSION dataset includes a predefined training-test split, the stratification criteria used are undocumented. To approximate the original criteria, the distribution of key attributes is analyzed across the original train and test sets, to get a better understanding of the baseline used.

To maintain comparability with the baseline, the original test set is preserved. The training set is re-split using various stratification configurations. All splits include the target labels, and additional attributes based on known representation disparities are incorporated. A purely random split serves as a control configuration.

Splits are generated using `sklearn.model_selection.train_test_split` with the `stratify` parameter. The general evaluation follows the procedure described in section 4.7.

5 Execution

Dies ist das Hauptkapitel Ihrer Arbeit! Hier wird die Umsetzung der eigenen Ideen und Konzepte (Kapitel 3) anhand der gewählten Methoden (Kapitel 4) beschrieben, inkl. der dabei aufgetretenen Schwierigkeiten und Einschränkungen. Die gewählten Methoden werden systematisch, konsistent und korrekt auf den Kontext der Arbeit angewendet. Die Bearbeitungs- bzw. Forschungsobjekte sind einheitlich benannt, im Kontext dargestellt und sinnvoll in die Arbeit integriert. Praxis- und Erfahrungswissen (z.B. aus Interviews) wird zur Validierung und Ergänzung der erarbeiteten Ergebnisse herangezogen.

5.1 Contextualization and Scope Definition

This section applies the information found during the literature review to the PASSION project using the method described in section 4.3. It also scopes what information can be assessed during this thesis and what should be passed on to the PASSION team.

5.1.1 Bias

5.1.2 Sensitive Features

Some of the listed features in Table 2.2 were also mentioned in the dermatology context and/or are included as metadata in the PASSION dataset. Therefore, potential biases associated with them should be evaluated in the PASSION model.

Since PASSION aims to improve classification of skin diseases based solely on image data without any metadata, it does not use these factors as features for training, except for characteristics that are implicitly visible in the images. This is primarily the *skin type* (including the undertone). More broadly defined, the *socioeconomic status* and *geographic location* can also be leaked to the model through the images, due to their impact on disease presentation and progression. Since the model can access these characteristics during training, they can introduce bias and should therefore be closely examined.

Age and *sex* are generally not visible in the images. Also, *socioeconomic status* and *geographic location* do not necessarily need to lead to visual effects. However, since they can influence disease prevalence and are prone to bias, the PASSION model should be evaluated for potential bias regarding these characteristics.

The potential impact of *ethnicity* and *disabilities* on visual presentation or prevalence of dermatological conditions has not been assessed in this thesis, due to time constraints. It is recommended that the PASSION team investigates these aspects further.

The other sensitive feature seem not to be further relevant for PASSION.

5.1.3 Fairness Metrics

This chapter focuses on those fairness metrics which are able to evaluate demographic fairness and are applicable to the dermatology context of PASSION. Those are mainly *equalized odds* by Hardt et al. (2016) and *subgroup fairness* by Kearns et al. (2018).

In the context of PASSION, fairness metrics which consider both true positives and false positives are particularly relevant. A *true positive* indicates that a disease was detected correctly, while a *false positive* corresponds to a diagnosis of a disease that is not actually present. Including false positives helps to identify cases where individuals from certain demographic groups may be unfairly more likely to receive unjustified diagnoses. This has also been indicated by Sabato et al. (2024).

From the listed group fairness metrics in Table 2.3, only equalized odds considers true and false positives, which should therefore be used for the evaluation of PASSION. A detailed explanation of equalized odds is provided in subsection 2.3.2.

Given the specific dermatology use case in the context of PASSION, it is not clear whether individual fairness metrics would be feasible to use. Certain metrics propose to change attributes. This approach is not feasible for the skin type which is passed on to the model implicitly through the picture. Therefore, this thesis focuses on the group fairness metrics for now.

Given the demographic focus of this study and the composition of the PASSION dataset, subgroup fairness is particularly important. Therefore, this thesis aims to incorporate equalized odds on subgroups as a core metric for evaluation.

5.1.3.1 Limitations of Fairness Evaluation with Equalized Odds for PASSION

Fairness metrics such as equalized odds are originally defined for binary classification problems, typically considering binary labels and binary demographic groups. As a result, fairness libraries like Fairlearn offer implementations of these fairness metrics only for binary classification tasks (contributors, n.d.). To evaluate fairness in multiclass settings using these libraries, certain considerations are required. This chapter introduces the two key challenges for the fairness evaluation of PASSION, handling multiclass labels and multiple subgroups.

Multiclass Labels

In binary settings, fairness can be evaluated through simple comparisons of false positive and false negative rates. However, in multiclass classification, fairness must account for the full structure of the confusion matrix. Sabato et al. (2024)

generalizes equalized odds to multiclass classification by defining: *”For each $y, z \in \mathcal{Y}$, the value of $\mathbb{P}[\hat{Y} = z \mid Y = y, G = g]$ is the same for all $g \in \mathcal{G}$.”*

In practice, this means the entire confusion matrix must be equal across groups to satisfy strict multiclass fairness under equalized odds (Sabato et al., 2024). The similar approach is purposed by Putzel and Lee (2022).

More relaxed versions of multiclass equalized odds have also been proposed in the literature, such as applying equalized odds per class. However, researchers argue that such relaxations may not be suitable in all contexts, especially when different types of errors carry different consequences (Sabato et al., 2024; Putzel & Lee, 2022).

For instance, when the type of misclassification matters, equality of error rates is essential to ensure fairness, as noted by Putzel and Lee (2022). Furthermore, as Sabato et al. (2024) explicitly states, a fair classifier in healthcare should avoid differences in diagnosis errors for specific diseases across subgroups, since misdiagnoses can lead to different treatment outcomes.

Therefore, in PASSION, the strict version of the multiclass equalized odds should be preferred. However, the code provided by Sabato et al. (2024) was not easy reusable, and there is no such version included in libraries like Fairlearn. Therefore, this thesis uses the more relaxed version, since this is implementable with Fairlearn and is still able to provide first insights for PASSION.

Non-Binary Sensitive Features

There can also be non-binary sensitive features leading to multiple subgroups. The original definition of equalized odds does not account for this complexity. To generalize fairness evaluation to such settings, a one-vs-rest strategy can be applied. In this approach, each group is individually compared against the rest of the population (Nezami et al., 2024).

5.1.3.2 Fairlearn Implementation and Interpretation of Equalized Odds

Fairlearn provides the functionality to calculate equalized Odds by calculating equalized odds difference (EOD) and equalized odds ratio (EOR) and the class `MetricFrame` for a disaggregated report. It allows for the calculation of performance metrics based on sensitive attributes and supports the configuration of aggregation functions for summarizing subgroup disparities (contributors, n.d.).

For the calculation of the metrics, Fairlearn provides multiple configuration options. In this thesis, the settings `agg="mean"` and `method="to_overall"` are particularly relevant. This configuration reports the average difference between each subgroup’s performance and the overall performance, for a given type of subgroups (e.g., all possible subgroups based on FST and sex).

While it is also possible to report the worst-case deviation instead of the mean, this thesis focuses on an initial fairness assessment of PASSION. Therefore, using the mean as an aggregate measure is considered sufficient. For a more critical or risk-focused analysis, worst-case metrics should also be considered.

When comparing models, additional aggregation is necessary because Fairlearn reports fairness metrics separately for each type of subgroup. To identify the fairest model based on aggregated statistics across all subgroups, the following indicators should be considered:

- **Lowest average and median EOD**: reflects strong overall fairness across subgroups.
- **Low standard deviation of EOD**: indicates consistent performance and minimal disparity among subgroups.
- **Lowest worst-case EOD**: captures the fairness for the most disadvantaged subgroup by highlighting the largest deviation.

These metrics were selected based on the principle that a lower EOD indicates higher fairness, as a difference of 0 represents perfect equalized odds. For EOR, the interpretation is inverted: a value closer to one signifies higher fairness, while lower values indicate greater disparity (contributors, n.d.).

5.1.4 Mitigation Methods

TODO: **write mitigation methods chapter**

5.2 PASSION Dataset Assessment

The practical analysis is conducted according to the methods outlined in section 4.4:

- **Metadata Completeness.** The available PASSION metadata listed in Table 2.1 is compared to the demographic factors which are relevant for bias detection. Missing attributes are listed in subsection 6.1.1.
  For certain attributes, the impact on dermatology specific use case is not entirely clear based on the literature review. For the attributes sex and age which are used in the PASSION dataset, the author of PASSION was contacted to provide more insight about their impact. This information was incorporated in the literature review.
  In order to provide the most complete view possible, all attributes which might have an impact are listed for the PASSION team to double-check with a dermatologist.
- **Presence of Proxy Variables.** Since the intended purpose of the variables are not mentioned in the paper, the analysis for proxy variables was more difficult then expected. The result is based on the sensitive features and biases mentioned in the literature.
  Also, what the country variable represents in PASSION is not entirely clear based on the documentation. To clarify its meaning, the main author of PASSION was contacted. For all variables which appear to potentially be used as a proxy variable, recommendations are provided for more precise alternatives for the PASSION team to check.

- **Representation of Main Groups.** Since there is no Jupyter Notebook script provided by PASSION to gather the proportions in depth, a python script is created to gather this data, what increased the time effort for the detailed analysis. The script is part of the newly created `evaluator` class and is meant to be executed standalone. It prints the distribution as absolute support and percentage for all values of the attributes country, sex, fitzpatrick, impetig, conditions_PASSION, and ageGroup. The age group contains the ages binned into 5 year intervals, like it has been done by Gottfrois et al. (2024) in their distribution analysis. Also, it saves the distribution in a csv and prints a plot per attribute. The comparison between the values is done manually for now, since there are not too many values.
  TODO: ensure to discuss the evaluator class beforehand somewhere and add command to command in readme(evaluator.run_split_distribution_evaluation)
- **Representation of Relevant Subgroups.** The demographic distribution figures of PASSION are briefly analyzed for an initial indication of the representation of age and sex.

5.3 Reproducing PASSION Results

While attempting to reproduce the results reported in the PASSION paper, some issues in the provided codebase had to be addressed. First, the metadata filenames referenced in the code were outdated, and the linkage between images and metadata records seemed to not fit the provided metadata files, preventing proper data loading. This was resolved using the same method as in the ”Linking CSV Data with Image Files” script included with the PASSION data analysis scripts, ensuring compatibility. After fixing the data linkage, the models for `conditions_PASSION` and `impetig` were trained, and the results were compared with those reported in the PASSION publication.

During the verification of group-level performance reproducibility, it was identified that the linkage between predictions and metadata was not functioning correctly in the evaluation pipeline. The original linkage used indices, which proved unreliable. To confirm the issue, the trained model was reloaded and the evaluation rerun. If group-level evaluation metrics changed despite identical model and data inputs, the linkage must be faulty.

To allow for the model reloading, the checkpoint handling had to be revised. The evaluation process was encapsulated within a separate `Evaluator` class to improve code modularity and separation of concerns.

The corrupted metadata linkage was resolved by adding the image filename into the dataloader, allowing the `Evaluator` to accurately link predictions to the correct metadata records.

These unanticipated code fixes required significant time, but they were essential for ensuring the validity of the analysis.

5.4 PASSION Baseline Fairness Assessment

5.4.1 Baseline Setup

This evaluation was conducted on the *conditions_PASSION* model. The binary *impetig* model was excluded due to the already high complexity and runtime demands of the multiclass setup.

The original PASSION model was trained using a *ResNet50* architecture. However, due to its long training and evaluation time, a smaller model version, *ResNet18*, was used for the experiments to enable faster iteration. To get insight in potential performance disparities based on this substitution, both models were evaluated using the same fairness assessment process as described in chapter 4.6 Fairness Assessment. This enabled a comparison to verify whether the smaller model produced comparable subgroup fairness insights and could be reliably used for the experimental phase. TODO: try to cite, or at least use protocols

To further improve runtime efficiency and flexibility during the experiments, the several modifications were made to the original pipeline and methodology:

- Temporarily enabled parallel data loading to accelerate experimentation (this change was later reverted for better reproducibility).
- Accelerated data loading by moving redundant checks out of a loop.
- Introducing the concept to check variants of a mitigation method without 5-fold cross validation to allow for faster iterations

5.4.2 Fairness Assessment Implementation

Fairness was assessed using *equalized odds* on sensitive subgroups defined by unique combinations of *FST, sex, age group, and country*, as introduced in previous chapters. The evaluation was implemented following the method described in section 4.6.

Considering the findings in subsection 5.1.3, Fairlearn methods where combined with custom implementation to implement the relaxed version of multiclass equalized odds. The final evaluation consists of several steps:

- **Data Aggregation:** Prediction results and metadata are linked and saved into a unified CSV, which can be used for manual inspection and is loaded on evaluation reruns.
- **General Performance:** Overall performance metrics are reported, as implemented by the PASSION team.
- **(Sub-)group Evaluation:** For each combination of sensitive attributes, performance and fairness metrics are computed.
- **Class-Level Fairness Metric Computation:** Using `MetricFrame` from Fairlearn, EOD and EOR are computed per class. Due to the binary limitation of Fairlearn's implementation, a one-vs-all strategy is applied to enable multiclass fairness evaluation.

- **Aggregation on Subgroup Level:** Class-level fairness metrics are further aggregated per subgroup using:
  - Worst-case
  - Mean
  - Best-case

  This aggregation approach is inspired by the *summary* aggregation for subgroup reporting for one class provided by Fairlearn (contributors, n.d.)
- **Aggregation on Model Level:** The subgroup level metrics are aggregated further, to report fairness across all subpopulations for easier model comparison, using:
  - Worst-case
  - Mean
  - Median
  - Best-case
  - Standard deviation

  This last step is done manually using an *Excel*-file so far.

To identify all privileged and underprivileged subgroups, comparisons of subgroup TPR and FPR against macro-averages of the same type of subgroups were conducted. The rates where computed based on the confusion matrices. A relaxed threshold of 0.2 was used to ignore slight differences in this initial fairness assessment. Subgroups with better-than-average TPR and lower-than-average FPR were marked *privileged*; the inverse as *underprivileged*. Borderline groups were labeled *unclear*, and those lacking support were tagged with *no support*. Those outputs were cross-validated against manual calculations and Fairlearn's outputs for correctness. TODO: cite / Add reference to methods for multiclass fairness.

For comparing models, the aggregated values have to be compared. Currently, this step is also covered in the mentioned Excel file.

5.5 Stratified Split Experiment

To analyze the original split, the script from section 5.2 was extended to evaluate attribute distributions across each subsets.

The following attribute combinations were used for stratification:

1. conditions_PASSION, impetig
2. conditions_PASSION, impetig, country
3. conditions_PASSION, impetig, fitzpatrick
4. conditions_PASSION, impetig, country, fitzpatrick
5. conditions_PASSION, impetig, country, fitzpatrick, sex
6. Random split without stratification

A key challenge was the presence of subgroups with single records, which hinder stratification since at least two samples per subgroup are required for even distribution. These single-record instances were handled in two ways:

- Strategy A: Assigning them to the training set, ensuring the model learns from all subgroups but excluding them from fairness evaluation.
- Strategy B: Assigning them to the validation set, allowing subgroup inclusion in fairness analysis but excluding them from model training.

Both strategies were applied to each split, resulting in 12 models. The seeds got fixed to remain compatibility. Unfortunately, the seed was mistakenly changed between generating the different strategies. Therefore, the models were evaluated per strategy, to avoid improper comparisons.

To evaluate fairness, the models were trained using PASSION's pipeline with each split configuration. The evaluation focused on fairness metrics alone, given that this was the primary objective. Final evaluations included both fairness and performance trade-offs using 5-fold cross-validation on the most promising splits.

5.5.1 Limitations

Fairness evaluation was done entirely based on EOD, as EOR reported mostly values close to zero across most subgroups. This trend is consistent over the models, rendering EOR uninformative for this experiment.

Furthermore, skewed subgroup distributions often led to extreme TPR and FPR values, especially in small subgroups. This heavily affected the resulting EODs. Future work should address this by collecting more subgroup-specific data.

Lastly, the evaluation was challenging due to the number of models to comare and the required manual intervention. The process in subsection 5.4.2 needs to be improved.

6 Evaluation and Validation

Auswertung und Interpretation der Ergebnisse. Nachweis, dass die Ziele erreicht wurden, oder warum welche nicht erreicht wurden. Die Ziele / Forschungsfragen sind dem Umfang der Arbeit entsprechend sehr klar abgegrenzt; sie sind präzise, überprüfbar und nach den Standards der Zielformulierung definiert. Die Zielerreichung wurde systematisch und korrekt validiert. Die Herleitung und Bedeutung der Ergebnisse, mögliche Varianten, Gütekriterien und eine Validierung allgemein werden nachvollziehbar diskutiert

6.1 PASSION Dataset Assessment

The PASSION dataset assessment results are described in this section. Overall, the dataset enables a foundational fairness analysis but does not support in-depth bias evaluation without augmentation or careful interpretation.

6.1.1 Metadata Completeness and Proxy Variables

Based on the literature regarding sensitive features and potential biases, sensitive metadata is available in the dataset, namingly FST, age, sex, and country. To obtain a feasible number of comparable subgroups, age can be grouped into age groups by using 5-year age groups, following the approach by Gottfrois et al. (2024). However, relevant metadata for a thorough fairness assessment and bias mitigation is missing. This limits what biases can be detected.

The missing metadata attributes are:

- socioeconomic status
- geographic location / residence of the patient
- (type of) the clinic and their medical focus
- image quality or other image related information such as the phone used, whether the image contains hair, and so on
- ethnicity (if it proves to have an impact on dermatology conditions)
- disabilities (if it proves to have an impact on dermatology conditions)

The variable *country* currently could theoretically serve as proxy variable for *geographic location*, which clinic the data is from and more broadly even for the *image quality*. It is not clear if those usages are intended. According to the literature review, this should be prevented. TODO: ensure that this is indeed written somewhere in the literature section Since the country only reflects the location of diagnosis, it is insufficient to determine the *geographic location* or residence of the patient. More precise data would be preferable for robust bias analysis. Since the data is gathered only from one clinic per country, this proxy variable usage is feasible for now. However, more clinics should be included into the data collection process to mitigate medical biases and ascertainment bias. Then, the clinic and some more data about it should be added to the dataset. The clinic again might be a proxy variable for the picture quality. If this information can be quantified in another way, e.g., the used phone and camera settings, that would further improve the dataset by tackling image biases. The country information can still be used in the fairness assessment to see if there are fairness differences in those populations. However, in order to clearly identify related biases, the suggested changes to the metadata would need to be introduced.

It is suggested to add the missing metadata attributes to the dataset. Given the sensitivity of those attributes, ethical considerations must be addressed before extending the dataset.

6.1.2 Demographic Representation

The demographic distribution in the PASSION dataset shows clear imbalances across several attributes. The data is available in appendix D PASSION Dataset Distribution Analysis.

To summarize:

- **Country.** The dataset is heavily skewed towards samples from Madagascar (59.59%), while Tanzania is significantly underrepresented (1.39%). This imbalance may introduce geographic or clinic-specific biases.
- **Sex.** Male patients are overrepresented (58.2%) compared to female patients (41.8%). No data is available for individuals of other sexes.
  This thesis did not explore whether other biological sex differences or gender-affirming hormone therapies have any impact on dermatological conditions, since the main focus for PASSION is on inclusion regarding skin type. However, for a complete fairness evaluation, this factors should be explored in the future.
- **FST.** The types III to VI are represented, with the distribution ranging from 21.42% (type III) to 29.4% (type IV). No data is available for type II and only one sample for type I. Given PASSION's focus on highly pigmented skin, this distribution is somewhat justified. However, it limits applicability to lighter skin tones and could impair model generalizability.
  An interesting future direction would be to combine PASSION with other dermatology datasets to evaluate fairness and performance across the full spectrum of FST. Moreover, due to historical underrepresentation of highly-pigmented skin in dermatology datasets, the performance on types V (25.89%) and VI (23.23%) should be examined in more detail, to see if their representation in the dataset must be addressed further. TODO: cite https://academic.oup.com/bjd/ar
  abstract/185/1/198/6600283?redirectedFrom=fulltext, already mention this

in dermatology bias section

- **Age Groups.** Children aged 0–9 account for over 40% of the dataset, whereas elderly patients (65+) are nearly absent. Although this skew reflects PASSION's focus on pediatric conditions, the lack of data for seniors may reduce fairness for those age groups.
  Nevertheless, PASSION's age-generalization experiments suggest that a model trained on primarily pediatric images might generalize reasonably well (Gottfrois et al., 2024).
- **Conditions.** The dataset is dominated by fungal infections (35.02%), followed by scabies (28.49%), and eczema (25.05%). Other conditions account only for 11.43%. Mo data is available for healthy skin.
  The ambiguous "other" category complicates fairness evaluations for specific conditions. Disaggregating this group into defined labels would improve clarity. Additionally, including healthy skin samples could reduce potential bias and enable better calibration of diagnostic models. TODO: find and add healthy-vs-disease bias here
- **Impetigo Indicator.** The impetigo label is present in only 11.6% of the cases, indicating class imbalance that may affect prediction reliability for this condition.

The Figure 6.1 illustrates the overrepresentation of male children, based on the figures presented by Gottfrois et al. (2024). There are also condition-specific differences in FST distribution. If this imbalances significantly affect model fairness, the dataset composition may need to be revised.

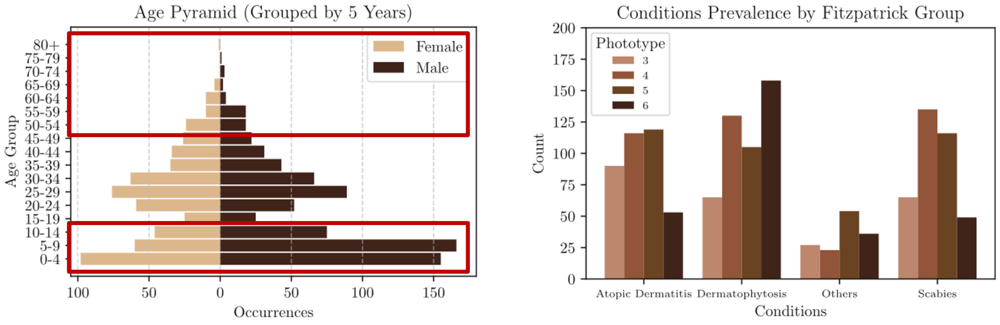


Fig. 1. Age distribution per gender. **Fig. 2.** Prevalence per FST.

Figure 6.1: PASSION dataset distributions by Gottfrois et al. (2024) - highlighting potential imbalances

These findings highlight representation disparities across several demographic and clinical factors. Such disparities should be accounted for during training and fairness evaluation, especially when assessing subgroup-specific performance.

It is important to note that the provided analysis only is a high-level overview at the group level. Detailed subgroup representation has not yet been assessed in details. Due to the time limits of this thesis, this was deferred in favor of executing the stratified split experiment.

To enable subgroup-level representation analysis, group-level dataset representation script should be extended accordingly. As the script output will increase substantially, manual comparison may become impractical. Therefore, automating the comparison and generating summaries of the largest disparities is recommended.

6.2 Reproducing PASSION Results

The overall model performance was consistent with the results reported in the PASSION paper.

However, the group-level performance results could not be reproduced. Multiple inference runs with the same model and dataset produced inconsistent results. Introducing metadata linkage via filenames resolved this issue and provided stable, reproducible results. This confirms a reliable association between predictions and metadata, which is critical for fairness analysis.

Currently, the checkpoint handling supports only evaluation. Additional adjustments are needed to fully support resumed training, particularly to ensure correct and reproducible handling of epochs and cross-validation folds.

Those changes will be contributed to the PASSION code base to make the reproduction easier for others.

While these extensive code improvements reduced the time available for fairness analysis, they are a critical enhancement to the robustness and usability of the PASSION evaluation.

6.3 PASSION Baseline Fairness Assessment

The baseline fairness performance of `ResNet50` and `ResNet18` was assessed. This evaluation serves as a reproducible reference against which the impact of fairness mitigation strategies can be compared.

- **Baseline ResNet50:** TODO: [Insert Equalized Odds Difference/Ratio metrics here per subgroup and model level aggregation]
- **Baseline ResNet18:** TODO: [Insert Equalized Odds Difference/Ratio metrics here per subgroup and model level aggregation]

Although minor performance differences between the two model versions were found (e.g., balanced accuracy of 0.70 for ResNet50 and 0.69 for ResNet18), the subgroup-level fairness trends proved largely consistent. Therefore, ResNet18 was considered a valid substitution model for experimentation in this thesis. TODO: add performance metrics as table

Further conclusions are summarized in the following subchapters.

6.3.1 Bias in the Baseline

TODO: link output Subgroup fairness results revealed some slight disparities, but still indicates a trend:

- **Sex:** The small model (`ResNet18`) showed no clear sex-related bias, whereas the larger model (`ResNet50`) exhibited a slight bias toward women, with higher TPR for female patients.
- **Skin Type (Fitzpatrick):** Both models consistently privileged Skin Types V and underprivileged Type VI. Notably, Skin Types III and VI showed different behavior across models, being more privileged in the big model.
- **Age:** The impact of age was relatively small overall. Nevertheless, age groups 0–14 and 25–29 were generally better off, whereas groups 20–24 and 30–69 were more often underprivileged. No samples from the 70+ group were available in the test data.
- **Intersectional Analysis (e.g., Sex × Age × Skin Type):** Subgroup-level analysis revealed distinct patterns, such as males aged 15–59 being consistently underprivileged across both models. Also, subgroups tend to have very low support, which makes the fairness analysis less stable.
- **Country:** Substantial differences emerged between countries. For example, Guinea performed better under the small model, while Malawi showed better results with the larger model. Tanzania remained underprivileged across both architectures.

Intersectional fairness issues also became apparent when combining protected attributes. For example, in the large model:

- **FST VI in Madagascar and Tanzania** performed particularly poorly.
- **Guinea with FST VI** still showed favorable outcomes, albeit slightly worse in ResNet50 compared to ResNet18. This indicates, that the country might impact the model's bias stronger than the skin type.

Overall, the clearest fairness disparities were observed in subgroups related to the attributes FST, sex, and country. Given PASSION's goal to mitigate bias against highly pigmented skin tones, fairness issues with FST VI are especially concerning. That some subgroups including FST VI and specific countries perform well indicates, that the bias could stem from other origins, such as the image quality or the process on how the data was gathered in those countries. While this analysis provides a first systematic fairness evaluation, deeper investigations are necessary, particularly into age-related intersectional effects. The provided scripts enable further detailed analysis and subgroup comparisons.

For further work, additional data collection efforts should prioritize Tanzania since the country is underrepresented which is consistently reflected in the model performance. Data quality or scarcity might be contributing to inconsistent results for this subgroup. Due to the sensitive medical nature of the images and personal limitations in handling such content, the images were not directly reviewed to support this hypothesis. It is, however, strongly recommended that the PASSION team conducts a thorough analysis of these cases.

6.3.2 Subgroup-Level Insights

Using the aggregation of class-level equalized odds metrics, the assessment revealed substantial variance across subgroups. Privileged and underprivileged subgroups are consistently identifiable.

While some groups showed stable behavior across classes, others shifted category depending on the evaluated class, which underlines the importance of per-class fairness computation in multiclass settings.

6.3.3 Pipeline Challenges

Several technical limitations impacted the reliability and completeness of the fairness assessment:

- Fairlearn's default multiclass handling is limited. To overcome this, a custom implementation was required, which introduces complexity and potential inconsistencies with the intended methodology of researchers.
- In the report part where subgroups are classified regarding privilege level, some subgroups were suppressed. This affects fairness analysis negatively. The comparison to the Fairlearn output revealed this issue. This proves that it is preferable to use well-established, tested libraries for whenever possible.
- Manual aggregation of subgroup-level to model-level metrics as well as the cross-model comparison is currently not automated, reducing reproducibility and increasing error risk.
- For model comparisons, the approach of Valentim et al. (2019) of creating fairness comparison rations could be used for the automated reporting.

Despite these challenges, the evaluation successfully surfaced subgroup disparities, supporting claims that fairness analysis on subgroups is important for reducing biases in dermatology models, including PASSION.

6.3.4 Aggregation Trade-offs

Aggregating fairness metrics at subgroup and model level provided helpful summaries but hide subgroup-specific effects. This must be considered when interpreting aggregated metrics.

The proposed aggregation strategy was implemented due to the absence of ready-to-use multiclass equalized odds metrics in Fairlearn or similar libraries. This illustrates the need for researchers to work together and implement suggested methodology improvements in the state of the art libraries.

6.4 Stratified Split Experiment

The evaluation of this experiment confirms that stratification strategies can influence fairness. The current findings suggest that including the attributes country and fitzpatrick in the stratification process can improve the fairness of models trained on the PASSION dataset. However, this improvement may come at the cost of overall model performance.

These results should be interpreted with caution, as the experiment faced several limitations. To achieve more statistically robust findings, additional data is needed. Further experiments should be conducted using the available codebase. Based on the results, the current PASSION split could likely be refined. Moreover, analyzing subgroup-specific performance may help guide future data collection efforts toward fairer outcomes.

6.4.1 Demographic Representation Accross Subsets

Distribution analysis of the original PASSION split shows that the distributions between the subsets are balanced for some attributes (e.g., country, conditions_PASSION), while others show notable discrepancies (fitzpatrick and sex). This suggests the original split may have used country and conditions_PASSION for stratification, possibly including impetig and ageGroup. TODO: link to appendix or github

Interestingly, the dataset shows male overrepresentation, especially in the training set, despite slight female bias in model performance (Table 6.1). Similarly, FST IV and V are overrepresented in training data (Table 6.2), possibly contributing to the observed bias. However, FST VI is evenly distributed accross the subsets, yet model performance remains poor. This suggests that data imbalance is not necessarily the sole cause of observed biases. However, a more detailed subgroup-level analysis across all subsets is still essential for a robust interpretation.

Set	**Female**	**Male**	**Total**
Training set	539 (40.74%)	784 (59.26%)	1323
Test set	152 (46.06%)	178 (53.94%)	330
Overall	691 (41.8%)	962 (58.2%)	1653

Table 6.1: PASSION Dataset: Sex distribution (train, test, overall).

TODO: consider to move tables in appendix

Set	**I**	**II**	**III**	**IV**	**V**	**VI**	**Total**
Training set	1 (0.08%)	–	275 (20.79%)	396 (29.93%)	344 (26.00%)	307 (23.20%)	1323
Test set	–	–	79 (23.94%)	90 (27.27%)	84 (25.45%)	77 (23.33%)	330
Overall	1 (0.06%)	–	354 (21.42%)	486 (29.40%)	428 (25.89%)	384 (23.23%)	1653

Table 6.2: PASSION Dataset: FST distribution (train, test, overall).

6.4.2 Initial Training

The fairness assessment of the initial model training as shown in Table 6.3 indicated that for strategy A, placing single records in training data, the configuration 4, using country and fitzpatrick, resulted in the fairest model overall, due to the analysis of reported EOD:

- Lowest average and median
- Fairly low standard deviation
- Moderate worst-case fairness

For strategy B (Table 6.4), placing singletons in validation data, the fairest model was achieved by stratifying only on the target labels, due to:

- Low average and median
- Lowest standard deviation
- Moderate worst-case fairness

Metric	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6
avg	0.53	0.55	0.56	0.48	0.54	0.55
best	0.03	0.03	0.04	0.05	0.01	0.08
worst	0.74	0.81	0.83	0.79	0.79	0.78
median	0.55	0.54	0.60	0.44	0.53	0.56
std. dev. sub pop.	0.22	0.23	0.25	0.23	0.25	0.22
std. dev. whole pop.	0.21	0.22	0.24	0.23	0.24	0.21

Table 6.3: Stratified Split: Fairness summary (seed 42, single-record training stratification).

Metric	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6
avg	0.51	0.55	0.57	0.55	0.55	0.49
best	0.02	0.04	0.03	0.03	0.05	0.03
worst	0.74	0.82	0.84	0.75	0.71	0.73
median	0.53	0.58	0.56	0.56	0.65	0.50
std. dev. sub pop.	0.18	0.25	0.23	0.20	0.19	0.22
std. dev. whole pop.	0.17	0.25	0.22	0.20	0.18	0.21

Table 6.4: Stratified Split: Fairness summary (seed 32, single-record validation stratification).

Random splits also performed surprisingly well. Using strategy B, even the highest fairness was achieved in terms of average and median, though with higher variance. Using strategy A, the lowest standard deviation was achieved, but the fairness was lower overall.

Due to these observations, and their overall well performance, splits 1, 4, and 6 were selected for 5-fold cross-validation.

6.4.3 Cross-Validation

Results of the 5-fold cross-validation step confirmed that including country and fitzpatrick in the stratification consistently resulted in the fairest models for both single-records-handling strategies (Table 6.5, Table 6.6).

Metric	Split 1	Split 4	Split 6
avg	0.54	0.50	0.48
best	0.03	0.03	0.03
worst	0.75	0.65	0.71
median	0.57	0.55	0.53
std. dev. sub pop.	0.20	0.16	0.20
std. dev. whole pop.	0.19	0.16	0.20

Table 6.5: Stratified Split: Fairness summary (5-fold CV, seed 32, validation stratification).

Metric	Split 1	Split 4	Split 6
avg	0.55	0.50	0.55
best	0.04	0.04	0.06
worst	0.82	0.77	0.77
median	0.54	0.51	0.58
std. dev. sub pop.	0.25	0.23	0.22
std. dev. whole pop.	0.24	0.22	0.22

Table 6.6: Stratified Split: Fairness summary (5-fold CV, seed 42, training stratification).

6.4.4 Baseline Comparison

Final evaluation on the original test set (Table 6.7) confirms that applying stratified splitting impacts model fairness, especially on subgroup levels. Stratifying also on country and fitzpatrick improved fairness, especially when single records are in the training set. For the other strategy, results are mixed, but also there, an impact is notable. Note the results are not directly comparable though due to different seeds.

Metric	Baseline	Strategy A	Strategy B
Overall			
avg	0.49	0.47	0.51
best	0.03	0.02	0.03
worst	0.73	0.75	0.74
median	0.54	0.51	0.54
std. dev. sub pop.	0.24	0.21	0.22
std. dev. whole pop.	0.23	0.21	0.22
Avg. Per Subgroup			
fitzpatrick	0.10	0.14	0.19
sex	0.03	0.02	0.03
ageGroup	0.46	0.43	0.54
country	0.34	0.36	0.33
fitzpatrick, sex	0.18	0.20	0.28
fitzpatrick, ageGroup	0.67	0.60	0.68
fitzpatrick, country	0.51	0.51	0.50
sex, ageGroup	0.56	0.53	0.62
sex, country	0.38	0.41	0.37
ageGroup, country	0.73	0.48	0.64
fitzpatrick, sex, ageGroup	0.70	0.75	0.74
fitzpatrick, sex, country	0.54	0.51	0.54
fitzpatrick, ageGroup, country	0.73	0.60	0.70
sex, ageGroup, country	0.73	0.69	0.74
fitzpatrick, sex, ageGroup, country	0.73	0.75	0.74

Table 6.7: Stratified Split: Fairness comparison: baseline vs. stratified variants.

While fairness improved in the stratified variants, this came with a noticeable drop in overall model performance (Table 6.8). This was somewhat expected, as the baseline used the full original training set, whereas the stratified variants employed an additional train-validation split, reducing the number of training samples. Both F1-score and balanced accuracy decreased compared to the baseline. Strategy B, in particular, exhibited the lowest performance across most metrics, likely due to having the smallest training set and lacking certain rare cases.

This illustrates the trade-off between fairness and predictive performance, which must be carefully managed in real-world applications.

Metric	**Baseline**	**Strategy A**	**Strategy B**
Accuracy	0.69	0.61	0.59
Macro F1	0.69	0.61	0.59
Weighted F1	0.69	0.62	0.59
Balanced Accuracy	0.69	0.62	0.60

Table 6.8: Stratified Split: Performance comparison: baseline vs. stratified variants.

7 Outlook

Reflexion der eigenen Arbeit, ungelöste Probleme, weitere Ideen. Die Ergebnisse und Empfehlungen schaffen einen konkreten Mehrwert für die Auftraggebenden. Einschränkungen und Grenzen werden kritisch diskutiert und die nächsten Schritte im Ausblick festgehalten, so dass die Ergebnisse direkt in der Praxis weiterverwendet und/oder angewendet werden können.

This chapter summarizes the concrete recommendations to overcome the limitations of the current work. This includes e.g., revising the metadata used in PASSION, and extending the analytical tools used.

It also provides ideas, such as adding more diverse data and combining PASSION with other dermatology datasets to improve bias detection and aim for a more complete dataset.

These measures aim to enhance the practical applicability of the results and support the development of fair, generalizable ML models in dermatology.

7.1 PASSION Dataset Improvements

To improve the fairness assessment capabilities of the PASSION dataset, the following dataset improvements are proposed:

- Include the missing metadata attributes identified in subsection 6.1.1 (e.g., socioeconomic status, clinic type, image quality) to enable a more comprehensive fairness evaluation. Ensure to assess the ethical implications before collecting such data.
  - Investigate whether *ethnicity* and *disabilities* influence the presentation or prevalence of dermatological conditions before adding them to the dataset.
- Clarify the intended purpose of the *country* variable, and replace or supplement it with more precise alternatives, as discussed in subsection 6.1.1.
- Refine the ”other” condition category by breaking it down into more specific labels to improve diagnostic granularity and fairness assessment per condition.
- Incorporate healthy skin samples into the dataset to allow for a more balanced classification task and to mitigate potential bias.
- Explore whether combining PASSION with other dermatology datasets enhances generalization across the full FST range.

7.2 Training Process Improvements

To enable full reproducibility and extensibility, further work should include:

- Finalizing checkpoint loading support for resumed training by correctly tracking and reloading epochs and folds.
- Incorporating automated tests to verify linkage integrity and model reproducibility.

7.3 Fairness Assessment Process Improvements

The measures to improve the fairness assessment process further are:

- Extend the existing dataset representation script, as described in section 5.2, to support subgroup-level analysis and automated comparison.
- Replace confusion-matrix-based fairness calculations with direct `MetricFrame`-based computation to streamline and unify the process.
- Improve subgroup handling in the fairness evaluation pipeline to include low-support groups more reliably.
- Automate all metric aggregation steps and document all assumptions clearly to enhance reproducibility.
- Introduce the model comparison ratio by Valentim et al. (2019).

7.4 Fairness Assessment Results Extension

TODO: @Proofreaders: habt ihr einen besseren Namen für dieses Kapitel? Es geht mir darum, dass weitere Analysen / Fairness assessments gemacht werden sollten
The existing fairness assessment results can be extended with those actions:

- Perform the representation analysis of relevant subgroups, as described in section 4.4 using the extended script, to determine whether observed unfairness stems from distribution imbalances at subgroup level.
- Evaluate model performance across FST types V and VI more closely and take measures if bias exist. TODO: check if this will still be needed
- Evaluate worst-case metrics and EOR during the fairness assessment and model comparisons as suggested in subsection 5.1.3. The metric computation is already included in the script but not yet useful due to missing data.
- Incorporate multiple training seeds for each experiment (also for the baseline) for drawing statistically valid conclusions about fairness across model variations and mitigation methods.

Implementing these measures will enhance the dataset's ability to support fair, robust, and generalizable ML models in dermatology.

7.5 Code Contribution

The code written during this thesis will be cleaned and provided as a pull request to the PASSION GitHub project, so that the team can us it for their future work.

8 currently working on

9 writing ongoing TODO REMOVE THIS CHAPTER

TODO: remove this chapter TODO: put this somewhere in the outlook Checkout this paper which suggest further methods and a flowchart to select the right fairness metric Barr et al. (2025)

10 Outlook PASSION Baseline Fairness Assessment

TODO: this must be summarized heavily In summary, the evaluation infrastructure and insights established in this thesis provide a meaningful first step toward robust fairness assessments in PASSION. With technical extensions and methodological refinements, it can evolve into a comprehensive toolset for bias detection and mitigation in dermatology and beyond.

The fairness evaluation presented in this thesis revealed critical insights into subgroup disparities and pipeline limitations within the PASSION project setup. While the implemented approach provided a reproducible and structured baseline, several directions for future work have emerged to improve the robustness, generalisability, and reproducibility of fairness assessments.

10.0.1 Metric Implementation and Evaluation Consistency

As discussed, the current fairness evaluation relies in part on confusion-matrix-based metric computations. Replacing these with direct use of Fairlearn's `MetricFrame` functionality across all fairness metrics would ensure a unified calculation standard and reduce implementation complexity. This would also increase consistency with other fairness research and future-proof the pipeline against methodological updates in fairness literature.

10.0.2 Improved Subgroup Handling

The suppression of low-support subgroups due to Fairlearn's internal handling and manual filtering mechanisms may bias the evaluation. This highlights the need for a more sophisticated strategy to include or resample underrepresented subgroups without compromising statistical reliability. Future efforts should aim to ensure that all relevant subgroups are fairly assessed, particularly in the context of medical datasets where representation imbalances are common.

10.0.3 Automation and Documentation of Aggregation

Currently, the aggregation of fairness metrics from class to subgroup to model level involves manual steps. These introduce the risk of inconsistency and reduce reproducibility. Automating these aggregation procedures with fully transparent logic and clearly documented assumptions would significantly improve the interpretability and replicability of fairness assessments. Where applicable, sensitivity analyses of aggregation strategies should be added to evaluate their impact.

10.0.4 Statistical Robustness through Multiple Seeds

One key limitation of this work is the single-seed evaluation due to time and computational constraints. In fairness research, especially with deep learning models, results can be sensitive to random initialization and data splits. As recommended in Valentim et al. (2019), future assessments should include multiple training and evaluation runs with different seeds to draw statistically significant conclusions about model fairness and mitigation impact. The current pipeline provides a foundation for such extensions.

10.0.5 Library and Methodology Development

The absence of well-established multiclass fairness implementations, particularly for metrics like equalized odds, required custom adaptations. This points to a broader need within the research community to extend existing libraries like Fairlearn to better support multiclass and subgroup-level analyses. Contributing the implemented methods and findings back to these open-source tools would benefit both the PASSION project and the wider fairness research community.

10.0.6 Broader Context and Integration

While this thesis focused on subgroup-level fairness analysis, future work could integrate fairness metrics into clinical utility assessments and decision-making contexts. This would enable a more holistic evaluation of trade-offs between fairness and clinical performance, which is essential for practical deployment in medical settings.

TODO: add somewhere that only the conditions classifier is used in this thesis, the impetig classifier should also be checked. (code is generalized, but must be tested)

11 Bias Evaluation using Equalized Odds in PASSION

TODO: fix this writing

11.1 Evaluation and Validation - baseline

A first analysis shows issues regarding xy in the big model and y in the small model.

There are some differences in the metrics per class based on the model size. To investigate each class individually would require more effort which should be done later. The provided scripts can be used to generate the required data. TODO: add specific info in the attachement Overall, the balanced accuracy for the small model = 0.69, big model = 0.7

small big Macro F1-Score: 0.69 0.71 Precision: 0.68 0.71 Sensitivity: 0.69 0.71 TODO: here were the scripts linked, check github

12 übergang zu bibliography

13 Bibliography

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July 16). A reductions approach to fair classification. https://doi.org/10.48550/arXiv.1803.02453

Alake, R. (2021, December 15). *How to read research papers: A pragmatic approach for ML practitioners* [NVIDIA technical blog]. Retrieved May 26, 2025, from https://developer.nvidia.com/blog/how-to-read-research-papers-a-pragmatic-approach-for-ml-practitioners/

Aleid, A. M., Nukaly, H. Y., Almunahi, L. K., Albwah, A. A., AL-Balawi, R. M. D., AlRashdi, M. H., Alkhars, O. A., Alrasheeday, A. M., Alshammari, B., Alabbasi, Y., & Al Mutair, A. (2024). Prevalence and socio-demographic and hygiene factors influencing impetigo in saudi arabian children: A cross-sectional investigation [Publisher: Dove Medical Press eprint: https://www.tandfonline.com *Clinical, Cosmetic and Investigational Dermatology*, *17*, 2635–2648. https://doi.org/10.2147/CCID.S472228

Baldé, B. (2023, April 14). *Why you should use stratified split* [Medium]. Retrieved April 14, 2025, from https://medium.com/@becaye-balde/why-you-should-use-stratified-split-bddb6dadd34e

Barr, C. J. S., Erdelyi, O., Docherty, P. D., & Grace, R. C. (2025, February 11). A review of fairness and a practical guide to selecting context-appropriate fairness metrics in machine learning. https://doi.org/10.48550/arXiv.2411.06624
Comment: 24 pages, 5 figures, 1 table.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*. Retrieved April 3, 2025, from https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
Mehrabi 20.

British Association of Dermatologists (BAD). (2021, July 7). *Lower socioeconomic status linked with more severe skin disease, including melanoma* [Bad patient hub] [Research was presented at the BAD's Annual Meeting.]. Retrieved February 17, 2025, from https://www.skinhealthinfo.org.uk/lower-socioeconomic-status-linked-with-more-severe-skin-disease-including-melanoma/

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification [ISSN: 2640-3498]. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–

91. Retrieved March 16, 2025, from https://proceedings.mlr.press/v81/buolamwini18a.html
Mehrabi 24, demographic (skin type and gender).

Chakraborty, A. (2024). Biases in dermatology: A primer [Publisher: Scientific Scholar]. *Indian J Dermatol Venereol Leprol*, *90*(2), 250–254. https://doi.org/10.25259/IJDVL_126_2023
0 citations (but from 2024), list of lots of biases.

Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024, July 1). Unmasking bias in AI: A systematic review of bias detection and mitigation strategies in electronic health record-based models. https://doi.org/10.48550/arXiv.2310.19917
Comment: Published in JAMIA Volume 31, Issue 5, May 2024.

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 339–348. https://doi.org/10.1145/3287560.3287594
Mehrabi 30.

contributors, F. (n.d.). *API docs — fairlearn 0.13.0.dev0 documentation*. Retrieved June 3, 2025, from https://fairlearn.org/main/api_reference/index.html

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. https://doi.org/10.1145/3097983.3098095
Mehrabi 41.

Delgado-Rodríguez, M., & Llorca, J. (2004). Bias [Publisher: BMJ Publishing Group Ltd Section: Continuing professional education]. *Journal of Epidemiology & Community Health*, *58*(8), 635–641. https://doi.org/10.1136/jech.2003.008466

Diaz, M., Lucke-Wold, B., Batchu, S., & Kleinberg, G. (2022). Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *3*, 42–47.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255
Mehrabi 48.

Farnadi, G., Babaki, B., & Getoor, L. (2018). Fairness in relational domains. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 108–114. https://doi.org/10.1145/3278721.3278733
Mehrabi 50.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., & Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *American Journal of Epidemiology*, *186*(9), 1026–1034. https://doi.org/10.1093/aje/kwx246
Mehrabi 54.

Gottfrois, P., Gröger, F., Andriambololoniaina, F. H., Amruthalingam, L., Gonzalez-Jimenez, A., Hsu, C., Kessy, A., Lionetti, S., Mavura, D., Ng'ambi, W., Ngongonda, D. F., Pouly, M., Rakotoarisaona, M. F., Rapelanoro Rabenja, F., Traoré, I., & Navarini, A. A. (2024). Passion for dermatology: Bridging the diversity gap with pigmented skin images from sub-saharan africa. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 703–712.

Grgic-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making
Mehrabi 61.

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining [Conference Name: IEEE Transactions on Knowledge and Data Engineering]. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459. https://doi.org/10.1109/TKDE.2012.72
Mehrabi 62.

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*. Retrieved March 16, 2025, from https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
Mehrabi 63.

Jayawickrama, T. D. (2021, February 1). *Community detection algorithms* [Medium]. Retrieved March 24, 2025, from https://medium.com/data-science/community-detection-algorithms-9bd8951e7dae

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness [ISSN: 2640-3498]. *Proceedings of the 35th International Conference on Machine Learning*, 2564–2572. Retrieved March 16, 2025, from https://proceedings.mlr.press/v80/kearns18a.html
Mehrabi 79.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 100–109. https://doi.org/10.1145/3287560.3287592
Mehrabi 80.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30*. Retrieved March 16, 2025, from https://proceedings.neurips.cc/paper_files/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
Mehrabi 87.

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., & Kohane, I. S. (2016). Genetic misdiagnoses and the potential for health disparities. *N Engl J Med*, *375*(7), 655–665. https://doi.org/10.1056/NEJMsa1507092
Mehrabi 98.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning [Publisher: ACMPUB27New York, NY, USA]. *ACM Computing Surveys (CSUR)*. https://doi.org/10.1145/3457607

Montoya, L. N., Roberts, J. S., & Hidalgo, B. S. (2025). Towards fairness in AI for melanoma detection: Systemic review and recommendations. In K. Arai (Ed.), *Advances in information and communication* (pp. 320–341). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-84460-7_21
2025.

Nezami, N., Haghighat, P., Gándara, D., & Anahideh, H. (2024). Assessing disparities in predictive modeling outcomes for college student success: The impact of imputation techniques on model performance and fairness [Number: 2 Publisher: Multidisciplinary Digital Publishing Institute]. *Education Sciences*, *14*(2), 136. https://doi.org/10.3390/educsci14020136

Pala, P., Bergler-Czop, B. S., & Gwiżdż, J. M. (2020). Teledermatology: Idea, benefits and risks of modern age – a systematic review based on melanoma. *Postepy Dermatol Alergol*, *37*(2), 159–167. https://doi.org/10.5114/ada.2020.94834

Petersen, K., Wohlin, C., & Baca, D. (2009). The waterfall model in large-scale development. Retrieved May 26, 2025, from https://urn.kb.se/resolve?urn=urn:nbn:se:bth-8073

Putzel, P., & Lee, S. (2022, January 12). Blackbox post-processing for multiclass fairness. https://doi.org/10.48550/arXiv.2201.04461

Romani, L., Whitfeld, M. J., Koroivueta, J., Kama, M., Wand, H., Tikoduadua, L., Tuicakau, M., Koroi, A., Ritova, R., Andrews, R., Kaldor, J. M., & Steer, A. C. (2017). The epidemiology of scabies and impetigo in relation to demographic and residential characteristics: Baseline findings from the skin health intervention fiji trial. *Am J Trop Med Hyg*, *97*(3), 845–850. https://doi.org/10.4269/ajtmh.16-0753

Sabato, S., Treister, E., & Yom-Tov, E. (2024, April 5). Fairness and unfairness in binary and multiclass classification: Quantifying, calculating, and bounding. https://doi.org/10.48550/arXiv.2206.03234

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017, November 22). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. https://doi.org/10.48550/arXiv.1711.08536

Mehrabi 142Comment: Presented at NIPS 2017 Workshop on Machine Learning for the Developing World.

Taylor, C. (2023, April 5). *Unbiased and biased estimators* [ThoughtCo] [Section: ThoughtCo]. Retrieved April 5, 2025, from https://www.thoughtco.com/what-is-an-unbiased-estimator-3126502

Valentim, I., Lourenço, N., & Antunes, N. (2019). The impact of data preparation on the fairness of software systems [ISSN: 2332-6549]. *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, 391–401. https://doi.org/10.1109/ISSRE.2019.00046

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776
Mehrabi 149.

Vickers, S. M., & Fouad, M. N. (2014). An overview of EMPaCT and fundamental issues affecting minority participation in cancer clinical trials. *Cancer*, *120*(0), 1087–1090. https://doi.org/10.1002/cncr.28569
Mehrabi 150.

Wang, T., & Wang, D. (2014). Why amazon's ratings might mislead you: The story of herding effects [Publisher: Mary Ann Liebert, Inc., publishers]. *Big Data*, *2*(4), 196–204. https://doi.org/10.1089/big.2014.0063
Mehrabi 151.

Young, A. T., Xiong, M., Pfau, J., Keiser, M. J., & Wei, M. L. (2020). Artificial intelligence in dermatology: A primer. *Journal of Investigative Dermatology*, *140*(8), 1504–1512. https://doi.org/10.1016/j.jid.2020.02.026
209 citations.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017, July 29). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. https://doi.org/10.48550/arXiv.1707.09457
Mehrabi 167 Comment: 11 pages, published in EMNLP 2017.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018, April 18). Gender bias in coreference resolution: Evaluation and debiasing methods. https://doi.org/10.48550/arXiv.1804.06876
Mehrabi 168, Comment: NAACL '18 Camera Ready.

Projektspezifisch können weitere Dokumentationsteile angefügt werden wie: Aufgabenstellung, Projektmanagement-Plan/Bericht, Testplan/Testbericht, Bedienungsanleitungen, Details zu Umfragen, detaillierte Anforderungslisten, Referenzen auf projektspezifische Daten in externen Entwicklungs- und Datenverwaltungstools etc.

TODO: fix appendices chapters, wtf TODO: also move the bibliography potentially after the appendices, idk

A PASSION Data Analysis Scripts

The PASSION team provides a Jupyter Notebook with code examples and analysis scripts. They are listed in Table A.1 together with their relevance to this thesis. The most relevant scripts are those related to demographic distributions of the chosen attributes, since they help identifying potential data imbalances. Scripts that lay the foundation for further analysis are somewhat relevant, while all other scripts are irrelevant for this thesis.

Script Title	**Description**	**Relevance - Reasoning**
Distribution of FSTs	Counts and visualizes the skin type distribution	**High** - Insight into demographic distributions
Regrouping Malawi and Tanzania to EAS	Data aggregation due to dataset size and geographical proximity	**Medium** - Might impact interpretation of the results of the following scripts
Linking CSV Data with Image Files	Mapping between data records and images.	**Medium** - Basis for other analyses
Extracting and Comparing Subject IDs	Dataset verification regarding completeness	**Low** - No insight in regards of demographic distribution
Conditions by Country	Correlation between clinical conditions and country	**Low** - The attribute *country* is out of scope of this thesis
Body Localizations by Conditions	Correlation between the condition and primarily affected body parts	**Low** - No insight in regards of demographic distribution
Impetigo Cases	Total count of impetigo cases and proportion to all cases	**Low** - No insight in regards of demographic distribution*

* Research is divided on which demographic factors influence the prevalence of impetigo (Romani et al., 2017; Aleid et al., 2024).

Table A.1: PASSION dataset - existing analysis scripts (Gottfrois et al., 2024)

B List of Biases

The biases are categorized and the relevance for PASSION is added to their chapter title in italic, e.g. *high*.

B.1 Category: Sampling Bias

Sampling biases occur when the process of collecting data results in samples that are not representative of the broader population. These biases affect the generalisability of machine learning models, especially in medical applications, where population diversity is crucial. According to Mehrabi et al. (2021), non-random or selective sampling can lead to serious consequences in terms of fairness and effectiveness of AI systems.

B.1.1 Sampling Bias, *high*

- **Definition:** Bias introduced through non-random sampling of subgroups, leading to poor generalisation.
- **Example:** An ML model trained predominantly on patients from urban hospitals may underperform for rural patients.
- **PASSION Relevance:** PASSION aims to address dermatologic sampling bias against highly pigmented skin, but if the included data is not truly representative across populations (e.g., over-representation of certain regions), this could still result in sampling bias (Mehrabi et al., 2021).
- **Mitigation Strategy:** Ensure a truly random and inclusive sampling strategy across geography, socioeconomic status, and skin types.

B.1.2 Selection Bias, *high*

- **Definition:** Bias arising when only a specific subset of the population is used, which is not representative.
- **Example:** Training a model only on adult data, when the target population includes children.
- **PASSION Relevance:** PASSION may suffer from selection bias if only data from severe dermatology cases in hospitals is used (Mester, 2022; Chakraborty, 2024).

- **Mitigation Strategy:** Include a broad variety of case severities and healthcare settings in the dataset.

B.1.3 Systematic Selection Bias, *high*

- **Definition:** A form of selection bias where chosen samples differ systematically from the general population.
- **Example:** Including only hospitalized patients in a dataset, while most cases are treated in outpatient settings.
- **PASSION Relevance:** If PASSION uses data only from dermatology centers treating severe cases, it introduces systematic selection bias (**c5**; **c6**; **c33**; Chakraborty, 2024).
- **Mitigation Strategy:** Include mild, moderate, and severe cases from various clinical settings.

B.1.4 Ascertainment Bias, *high*

- **Definition:** A systematic distortion arising from the method by which participants or data are selected for inclusion.
- **Example:** Studies on STD prevalence conducted only in public clinics may overlook patients from higher-income backgrounds who go to private practitioners.
- **PASSION Relevance:** If PASSION's dataset is composed mostly of patients from certain types of clinics, it may not generalise well to other socioeconomic groups (**c5**; Chakraborty, 2024).
- **Mitigation Strategy:** Ensure that data is collected from a diverse range of sources, including both public and private healthcare facilities.

B.1.5 Availability Bias, *high*

- **Definition:** Overreliance on easily accessible data rather than the most representative data.
- **Example:** Using only online available datasets for skin conditions may underrepresent rare diseases.
- **PASSION Relevance:** PASSION inherits availability bias by relying on FST scale–labelled datasets, which may not fully reflect global skin tone diversity (**c9**; **c10**; Chakraborty, 2024).
- **Mitigation Strategy:** Actively seek underrepresented data sources, especially for less common or less documented skin types.

B.1.6 Survivorship Bias, *medium*

- **Definition:** Only using data from ”survivors”, i.e., subjects that make it through a certain threshold or are retained in the dataset, ignoring those who were lost earlier.

- **Example:** Evaluating the success of a treatment based only on patients who completed it, ignoring those who dropped out due to side effects.
- **PASSION Relevance:** If certain dermatology diseases are lethal or if the dataset excludes patients unable to attend the centers involved in PASSION, survivorship bias may be present (Mester, 2022).
- **Mitigation Strategy:** Account for dropout rates and include cases from a wide range of medical access points.

B.2 Category: Representation Biases

Representation biases occur when a sample used to train or evaluate a machine learning model fails to adequately reflect the diversity of the target population. These biases can lead to underperformance for certain subgroups and may negatively impact the fairness and accuracy of a model in real-world applications. In the context of dermatology, these biases could result in skin diseases being underrepresented or misclassified in specific demographic groups, leading to poorer diagnostic outcomes for those populations.

B.2.1 Representation Bias, *high*

- **Definition:** Representation bias arises when the sample used to train a model does not adequately represent all subgroups of the target population, leading to missing or misrepresented characteristics in the data.
- **Example:** If a skin disease detection model is trained predominantly on skin types I-IV, it may struggle to accurately diagnose conditions in individuals with darker skin tones (FST skin types V-VI).
- **PASSION Relevance:** PASSION attempts to mitigate representation bias by including more FST skin types, but challenges may still exist. The dataset could still lack full representation of all diverse skin conditions and demographic factors, leading to potential misdiagnoses or underperformance for specific subgroups.
- **Mitigation Strategy:** A potential mitigation strategy could involve ensuring a more balanced representation of FST skin types, including rare and diverse skin conditions, and periodically reassessing the dataset to ensure comprehensive inclusion of all skin types across various demographics.

B.2.2 Population Bias, *medium*

- **Definition:** Population bias occurs when the sample's demographic characteristics (such as age, gender, or ethnicity) do not align with the target population, leading to non-representative data.
- **Example:** If a dataset is predominantly comprised of one ethnic group, a model trained on this data may not generalize well to other ethnic groups,

especially if the manifestation of skin diseases varies across ethnicities.

- **PASSION Relevance:** PASSION might be impacted by population bias if it is insufficiently diverse in terms of patient demographics (e.g., ethnicity, age). The dataset needs to ensure that skin diseases are accurately represented across different population groups to avoid skewing results and compromising diagnostic accuracy.
- **Mitigation Strategy:** A mitigation strategy could involve collecting data from diverse populations and ensuring the dataset reflects the target population's demographic diversity, particularly for ethnicities and age groups that may exhibit different disease manifestations.

B.2.3 Aggregation Bias, *high*

- **Definition:** Aggregation bias occurs when conclusions drawn from the entire population do not apply to individual subgroups, leading to incorrect or generalized assumptions. This bias arises when significant differences between subgroups (such as gender or ethnicity) are not properly accounted for.
- **Example:** A diagnostic model trained on a heterogeneous dataset might fail to capture how skin diseases manifest differently across genders or ethnic groups, potentially leading to misdiagnosis or unequal treatment recommendations.
- **PASSION Relevance:** Aggregation bias is a significant concern in PASSION, particularly since skin diseases can manifest differently across ethnicities, genders, or genetic backgrounds. The model needs to account for these variations to avoid generalized conclusions that might harm certain subgroups.
- **Mitigation Strategy:** To mitigate aggregation bias, the model should incorporate subgroup-specific data and analysis, ensuring that disease manifestations are correctly accounted for and tailored to different demographic characteristics.

B.2.4 Simpson's Paradox, *medium*

- **Definition:** Simpson's Paradox is a form of aggregation bias where trends that appear in aggregated data may reverse when the data is disaggregated into subgroups. This paradox can lead to misleading conclusions if not properly addressed.
- **Example:** A dataset may show that skin disease detection is more accurate overall for a specific demographic group, but when the data is broken down by age or skin type, the trend reverses for certain subgroups.
- **PASSION Relevance:** Simpson's Paradox could be an issue in PASSION if aggregated data from different subgroups results in misleading conclusions. For example, overall accuracy may appear high, but specific skin conditions

in certain ethnicities or age groups could have lower accuracy when analyzed separately.

- **Mitigation Strategy:** A mitigation strategy would involve analyzing data at both the aggregated and disaggregated levels, ensuring that subgroup-specific trends are considered to avoid false conclusions or the reversal of apparent associations.

B.3 Category: Measurement Biases

Measurement biases occur when the process of choosing, using, or measuring features leads to inaccurate or misleading results. These biases can emerge from various sources such as mismeasured variables, subconscious expectations of researchers, or inconsistencies in human annotation, and they can significantly affect the reliability of the dataset.

B.3.1 Measurement Bias, *high*

- **Definition:** Measurement bias occurs when features or variables are inaccurately measured or selected, leading to incorrect interpretations of the outcome.
- **Example:** If a proxy variable, such as country of origin, is used to infer ethnicity or genetic background, it could lead to misinterpretation of the data. For instance, the country of origin may not directly correlate with ethnic background, potentially skewing results in genetic or disease research (Mehrabi et al., 2021).
- **PASSION Relevance:** In the context of the PASSION dataset, measurement bias could arise if country of origin is misused as a proxy for ethnicity, which is not directly related to genetic predispositions or skin conditions. This could result in misleading conclusions about skin diseases across different demographic groups, potentially amplifying health disparities.
- **Mitigation Strategy:** To mitigate measurement bias in PASSION, careful consideration should be given to the choice of features used in the dataset. Avoiding proxy variables such as country of origin to infer ethnicity and instead focusing on genetically relevant factors could improve the accuracy of the data and its interpretation.

B.3.2 Observer Bias, *medium*

- **Definition:** Observer bias occurs when researchers or testers influence the results by projecting their expectations or perceptions onto the data collection process, or when different observers report the same observation differently.
- **Example:** A researcher may subconsciously interpret certain skin disease symptoms differently based on their own expectations or biases, leading to

inconsistent data collection or interpretation (Mester, 2022).

- **PASSION Relevance:** In PASSION, observer bias could affect the consistency and reliability of skin disease annotations. For example, a researcher might influence how they categorize or diagnose certain skin diseases based on their personal biases or experience. This could lead to inaccurate classifications, particularly for diseases that are subjective in appearance.
- **Mitigation Strategy:** To address observer bias in PASSION, standardized training for annotators and a clear, objective set of criteria for diagnosis should be implemented. Additionally, using multiple annotators and cross-checking results can help reduce the impact of individual biases.

B.3.3 Annotator Bias, *high*

- **Definition:** Annotator bias is a form of observer bias where human annotators are influenced by personal background, expectations, or external factors, which can lead to inconsistent or skewed labeling of data (Montoya et al., 2025).
- **Example:** If an annotator is more likely to label a darker skin tone as ”severe” or ”critical” due to personal or cultural biases, this can introduce inaccuracies in the dataset, which may not be representative of the actual severity of the condition.
- **PASSION Relevance:** In PASSION, annotator bias could particularly affect the labeling of skin tones, which are highly subjective and dependent on individual perception. This bias could lead to inconsistent classifications of skin conditions across different demographic groups, which is critical when assessing dermatological diseases in a diverse population.
- **Mitigation Strategy:** To reduce annotator bias in PASSION, a diverse team of annotators should be trained to recognize and overcome their personal biases. Additionally, the annotation process should be regularly audited to ensure consistency, and the use of automated tools for initial labeling could provide more objectivity in the process.

B.3.4 Recall Bias, *medium*

- **Definition:** Recall bias occurs when individuals do not accurately remember or report information due to selective memory, which can lead to misinterpretations or inaccurate conclusions in data analysis (Mester, 2022; Chakraborty, 2024).
- **Example:** If patients are asked to recall past skin conditions or treatments, they may forget important details, leading to inaccurate reporting in the dataset. This could affect the analysis of how different skin diseases develop or respond to treatments.
- **PASSION Relevance:** Recall bias may not be directly relevant in the context of PASSION since the dataset appears to rely on clinical observations and annotations rather than patient-reported data. However, if there is any

patient input, such as in follow-up surveys or self-reported symptoms, recall bias could still influence the dataset.

- **Mitigation Strategy:** To mitigate recall bias, it would be important to gather more objective data through clinical observations or imaging, and ensure that patient self-reports are validated through corroborating medical records or consistent follow-ups.

Potential Biases in PASSION Measurement Bias: Country of origin should not be used as a proxy for ethnicity in the PASSION dataset, as it may not be directly related to genetic or disease factors. Additionally, annotator bias regarding skin tone labeling has been investigated in recent studies and should be addressed in PASSION's annotation process (Montoya et al., 2025).

B.4 Category: Research Biases

Research biases refer to the ways in which researchers' decisions, intentions, and contexts influence the outcomes of their studies, potentially introducing systematic errors that may affect the validity or generalizability of the findings.

B.4.1 Funding / Sponsorship bias, *medium*

- **Definition:** Funding or sponsorship bias occurs when research findings are consciously or unconsciously influenced by the expectations or interests of the study's financial backers. This can lead to findings that favor the sponsor's interests.
- **Example:** A dermatology study funded by a pharmaceutical company that produces skin disease treatment medications may emphasize the effectiveness of the company's products, even if there is no strong evidence supporting their superiority.
- **PASSION Relevance:** Funding bias could affect the PASSION dataset if the research or data collection process were influenced by sponsors or stakeholders with vested interests in certain outcomes. While this bias is not explicitly mentioned in PASSION, it is important for future studies to ensure that funding sources do not shape data interpretation or collection in a way that would lead to skewed or misleading results.
- **Mitigation Strategy:** To mitigate this, independent funding sources or transparent funding disclosure practices should be implemented. Additionally, external audits or independent validation of the findings can help prevent undue influence from sponsors.

B.4.2 Data dredging bias, *low*

- **Definition:** Data dredging bias arises when researchers deliberately select statistical methods or models that lead to specific p-values or results, potentially making their hypothesis appear more likely to be true than it actually is.

- **Example:** A researcher testing multiple variables in a dataset might select those combinations that yield the most statistically significant results, even if the relationships between the variables were not hypothesized initially.
- **PASSION Relevance:** Given that PASSION is a large dermatology dataset, it could be vulnerable to data dredging if analysts test many variables or relationships without pre-specified hypotheses. This could lead to spurious findings or models that do not generalize well to new data.
- **Mitigation Strategy:** To avoid data dredging, a clear and well-defined hypothesis should be established before conducting any statistical tests. Additionally, cross-validation techniques and reporting of all tested models can ensure transparency in the research process.

B.4.3 Hypothetical bias, *not applicable*

- **Definition:** Hypothetical bias occurs when responses to hypothetical questions do not reflect real-world behavior or preferences.
- **Example:** Asking participants how likely they would be to adopt a particular skincare treatment, without actually testing their behavior in real-world settings.
- **PASSION Relevance:** This bias is not applicable to the PASSION dataset, as the dataset does not involve hypothetical scenarios or self-reported intentions. The dataset primarily contains real-world medical data related to dermatology, which does not rely on participant speculation or hypothetical responses.
- **Mitigation Strategy:** Since this bias is not relevant to PASSION, no specific mitigation strategy is necessary.

Potential Biases in PASSION

Since the PASSION dataset is already published, the research biases might already be introduced. It is not feasible during the duration of this thesis to make an evaluation on those biases. Instead, I would recommend the PASSION team and researchers in general to check the list above carefully and take measures against them. Maybe, an external evaluation could help to detect and prevent those biases even better.

B.5 Category: Feature Representation Biases

Feature representation biases occur when the features or variables used in a model do not adequately capture the complexity of the problem or reflect all relevant aspects of the data, potentially leading to biased or incomplete predictions.

B.5.1 Omitted Variable Bias, *high*

- **Definition:** Omitted variable bias arises when key variables are left out of a model, causing the model to be unprepared to account for certain aspects of the data and potentially leading to biased or inaccurate predictions.
- **Example:** If a dermatology model only includes skin condition data but omits important demographic information such as ethnicity or age, it may fail to identify or misinterpret certain patterns in the data.
- **PASSION Relevance:** The PASSION dataset has an omission of ethnicity as a feature, which could lead to biased results. Certain skin diseases and their manifestation can vary significantly across different ethnic groups. Without this variable, the model may fail to capture important differences in the data, leading to inaccurate predictions or generalizations.
- **Mitigation Strategy:** To address this bias, it is important to include a comprehensive set of features, such as ethnicity, age, gender, and other demographic factors, which could help the model better account for variations in skin conditions across different populations.

B.5.2 Collider Bias, *medium*

- **Definition:** Collider bias occurs when two variables influence a common third variable (the collider variable), and the analysis restricts sampling based on this collider, leading to a distorted or biased relationship between the variables.
- **Example:** In the case of skin disease models, if researchers only include patients who seek treatment for a specific skin condition (the collider), this may limit the analysis to a non-representative sample, potentially distorting the relationship between disease characteristics and other factors.
- **PASSION Relevance:** Although no specific collider bias has been identified in PASSION, it is important to consider that factors like patient willingness to seek treatment or the specific type of skin disease could act as collider variables. Restricting the dataset based on these factors might create a biased representation of the population.
- **Mitigation Strategy:** To reduce collider bias, it is important to ensure that the sample is as representative as possible of the broader population. Researchers should avoid restrictions that could inadvertently create a non-representative dataset and be mindful of how their sampling methods may introduce bias.

Potential Biases in PASSION The PASSION dataset may suffer from omitted variable bias, particularly with the lack of ethnicity data, which can affect the fairness and accuracy of dermatology models. Collider bias could also emerge depending on how the dataset is sampled or restricted based on treatment-seeking behavior or disease severity. It is important for researchers to monitor for these biases and take steps to mitigate them by ensuring that data collection and sampling strategies are inclusive and comprehensive.

B.6 Category: Imaging Biases

Imaging biases refer to the influence that technical variations, environmental factors, and other visual elements have on image-based classification systems. These biases can arise from issues such as the quality of the image, artifacts present in the image, or the field of view captured, which can all influence the performance of machine learning models.

B.6.1 Image Quality Bias, *high*

- **Definition:** Image quality bias occurs when the quality of an image—such as the zoom level, focus, or lighting—affects how a machine learning model classifies or diagnoses the image. Poor image quality can lead to misclassification or lower prediction accuracy.
- **Example:** If a dermatologist captures an image with insufficient lighting or poor focus, the model may struggle to identify skin conditions like melanomas, potentially leading to a misdiagnosis.
- **PASSION Relevance:** In the PASSION dataset, variations in image quality could lead to biased predictions. For instance, images captured under different lighting conditions or at varying zoom levels might cause the model to overfit to certain image qualities, mistaking them for certain conditions. This could reduce the model's generalizability to diverse real-world conditions.
- **Mitigation Strategy:** To mitigate image quality bias, it is essential to standardize image acquisition protocols and pre-process images to normalize variations in quality. Implementing techniques like image enhancement and quality control during data collection could help improve model performance.

B.6.2 Visual Artifact Bias, *high*

- **Definition:** Visual artifact bias arises from artifacts in dermatology images, such as hair, surgical ink markings, or other extraneous elements that could interfere with accurate classification of skin diseases.
- **Example:** A photograph of a skin lesion may contain hair or tattoos from previous medical procedures, making it more difficult for the model to identify the skin condition correctly.
- **PASSION Relevance:** The PASSION dataset may include dermatology images with artifacts like surgical markings or hair, which could confuse the model into associating these artifacts with the presence of a skin disease. This could lead to incorrect predictions, especially if the model cannot differentiate between the artifact and the actual lesion.

- **Mitigation Strategy:** To reduce visual artifact bias, it is important to implement preprocessing steps that remove or mask artifacts in images. This could involve techniques such as hair removal or the use of clean, artifact-free image samples for training.

B.6.3 Field of View Bias, *high*

- **Definition:** Field of view bias occurs when the portion of the body or skin that is captured in an image is limited, affecting how well a model can classify a skin condition. Different angles, distances, or body parts in the view may lead to different prediction results.
- **Example:** If only a small portion of a skin lesion is captured in the image (e.g., just the edge of a mole), the model may miss critical features needed to correctly identify melanoma or other conditions.
- **PASSION Relevance:** In the PASSION dataset, field of view bias could emerge if certain lesions are captured from angles or in parts of the body that limit the information available for accurate classification. This could result in the model underperforming on images that are not representative of common views of skin conditions.
- **Mitigation Strategy:** To address field of view bias, the dataset should ensure that images are captured from standardized and consistent angles or distances. Augmenting the dataset with a variety of views from multiple angles could help improve the model's ability to generalize to unseen cases.

Potential Biases in PASSION The PASSION model could learn to associate unrelated visual effects, hair, body parts, or image quality with a disease, which could impact its performance. Ensuring standardized image acquisition methods and removing artifacts could mitigate some of these biases.

B.7 Category: Medical Biases

Medical biases are specific to healthcare-related machine learning applications and can have direct implications for diagnosis, treatment, and patient outcomes. These biases often arise from the healthcare system's structure and can lead to distorted or inaccurate predictions based on incomplete or unrepresentative data.

B.7.1 Berkesonian Bias, *medium*

- **Definition:** Berkesonian bias occurs in hospital-based studies when certain factors (such as disease severity or risk factors) influence whether patients seek treatment or are hospitalized. This can distort the relationship between variables due to the study population being unrepresentative of the general population.

- **Example:** In a study focusing on skin diseases, if only patients who sought care for severe conditions are included, the relationship between disease severity and other factors could be overstated, leading to inaccurate conclusions.
- **PASSION Relevance:** The PASSION dataset could be influenced by Berkesonian bias if the images are sourced only from patients who visited certain hospitals or dermatologists, potentially skewing the representation of less severe or untreated conditions. This could limit the model's generalization to populations with different healthcare access.
- **Mitigation Strategy:** To mitigate Berkesonian bias, it is important to include a diverse set of patients from multiple sources, including both hospital and non-hospital populations, ensuring a more representative dataset.

B.7.2 Informed Presence Bias, *medium*

- **Definition:** Informed presence bias occurs when individuals who seek medical care are more likely to be screened for other diseases. This bias can result in misleading interpretations of the relationships between diseases.
- **Example:** A person who is already being treated for one skin condition might also be screened for other conditions, leading to a misinterpretation of comorbidities or a false relationship between conditions.
- **PASSION Relevance:** In the PASSION context, informed presence bias could affect correlations between different skin diseases. If patients with certain conditions are more likely to seek treatment, the model might overestimate the likelihood of co-occurrence between those conditions.
- **Mitigation Strategy:** To reduce informed presence bias, the model should account for patients with varying levels of care-seeking behavior and ensure that both treated and untreated conditions are represented in the dataset.

B.7.3 Diagnostic Access Bias, *medium*

- **Definition:** Diagnostic access bias occurs when individuals in certain geographical locations have better access to medical care, leading to earlier diagnosis and potentially higher disease prevalence in those regions.
- **Example:** Patients in urban areas with better healthcare access may receive earlier diagnoses of skin conditions like melanoma, while those in rural or underserved areas may have their conditions diagnosed at a later stage.
- **PASSION Relevance:** PASSION attempts to address diagnostic access bias by including samples from later stages of diseases. However, it could still be relevant if the dataset over-represents well-diagnosed cases from areas with better healthcare access, skewing the distribution of disease stages.
- **Mitigation Strategy:** To address diagnostic access bias, it is important to ensure that the dataset includes a diverse range of geographical locations and healthcare access levels, including both early and late-stage conditions.

B.7.4 Diagnostic Reference Test Bias, *medium*

- **Definition:** Diagnostic reference test bias occurs when not all individuals in a study receive the same reference test, leading to discrepancies in diagnoses.
- **Example:** Inconsistent use of reference tests across different hospitals or dermatologists may result in different diagnoses for the same patient, causing confusion and inconsistency in the results.
- **PASSION Relevance:** Depending on how dermatologists work in the PASSION dataset, diagnostic reference test bias could be present. If different diagnostic methods or reference tests are used, the model may learn to associate certain diagnostic practices with specific diseases, rather than the diseases themselves.
- **Mitigation Strategy:** To mitigate diagnostic reference test bias, it is important to standardize the diagnostic processes across different healthcare settings and ensure consistent use of reference tests when collecting data.

Potential Biases in PASSION Some of the medical biases that could impact PASSION include Berkesonian bias, informed presence bias, diagnostic access bias, and diagnostic reference test bias. Each of these could influence how the model generalizes to real-world populations. Addressing these biases requires careful consideration of the dataset's diversity and the standardization of diagnostic practices across different settings.

B.8 Category: Temporal Biases

Temporal biases arise due to differences in populations and their behavior over time. These biases can manifest when studying the progression of diseases or tracking changes in populations over time. In studies where data are collected over extended periods, temporal biases can affect the accuracy and generalizability of the results.

B.8.1 Longitudinal Data Fallacy, *not applicable*

- **Definition:** Longitudinal data fallacy refers to the misinterpretation or improper use of data collected over time, often caused by overlooking important variables or assuming temporal relationships without proper evidence.
- **Example:** A study might incorrectly assume that a disease progression observed over a period directly results from the treatment being applied, while other confounding factors may also play a role.
- **PASSION Relevance:** Temporal biases such as longitudinal data fallacy do not apply to the PASSION dataset, as it does not track disease progression over time but instead consists of static images that are not connected to temporal data.

- **Mitigation Strategy:** Since the PASSION dataset does not involve longitudinal data, no mitigation strategy is necessary for this particular bias.

B.8.2 Chronological Bias, *not applicable*

- **Definition:** Chronological bias occurs when the timing of data collection influences the results or introduces errors, often due to the use of data collected at different time points that may not be representative of the population or phenomenon being studied.
- **Example:** If a medical dataset includes only images collected from patients in a certain time period where a specific treatment was more commonly used, the findings might be skewed to reflect outcomes that are not generalizable to other time periods.
- **PASSION Relevance:** Chronological bias is irrelevant to the PASSION dataset as it consists of static images of skin diseases, without any temporal association or tracking of disease progression.
- **Mitigation Strategy:** Since PASSION does not contain temporal data, no mitigation strategy is needed for chronological bias in this case.

B.8.3 Immortal Time Bias, *not applicable*

- **Definition:** Immortal time bias refers to a situation where the period of time during which an event could have occurred is misclassified, leading to incorrect conclusions, typically when patients are erroneously considered ”at risk” for an event for a period in which the event could not have occurred.
- **Example:** A study that tracks patients who have received a specific treatment might misclassify the time between treatment and disease progression as time ”at risk,” even though the patients were not at risk during the follow-up period.
- **PASSION Relevance:** Immortal time bias does not apply to PASSION, as the dataset does not track time or disease progression and focuses on static images of skin diseases, eliminating the possibility of immortal time bias.
- **Mitigation Strategy:** No mitigation strategy is necessary for immortal time bias in PASSION, as it is not a relevant concern for the dataset.

B.9 Category: Algorithmic Biases

When an algorithm adds biases to unbiased input data, it is referred to as **Algorithmic Bias** (Baeza-Yates, 2018). This can arise due to various algorithmic design choices such as optimization functions, regularizations, and statistically biased estimators (Danks & London, 2017).

B.9.1 User Algorithm Interaction Biases, *high*

- **Definition:** User interaction biases arise when the user interface or user behavior influences the way an algorithm behaves, potentially introducing bias. This can occur when the user interface encourages specific actions or when users impose their own biases during interaction. (Baeza-Yates, 2018)
- **Example:** A user interacting with a teledermatology system might overrely on certain image features, skewing the algorithm's assessment or recommendation of treatment. For instance, if a teledermatology app visually emphasizes certain markers that are less important clinically, users may begin to prioritize those markers, which could distort the results the algorithm provides. Lerman and Hogg (2014) and Mehrabi et al. (2021)
- **PASSION Relevance:** In the PASSION project, user interaction biases could emerge as teledermatology platforms become more publicly available. As users interact with the system, they may unintentionally influence the algorithm's output, leading to biased diagnosis or treatment recommendations, particularly if the user interface highlights or prioritizes certain image features over others.
- **Mitigation Strategy:** To mitigate this bias, a careful evaluation of the user interface design is crucial. Ensuring that no unintended prioritization of image features occurs and that the interface does not suggest biases in how users should interact with the system would help. Additionally, the algorithm should be tested with diverse user interactions to ensure its robustness.

B.9.2 Emergent Bias, *high*

- **Definition:** Emergent bias occurs when changes in the population interacting with an algorithm cause shifts in how the algorithm behaves over time. These changes are not anticipated during the design phase and may appear after the algorithm is deployed. Emergent bias is especially common in user interfaces as they evolve with user behavior (Friedman & Nissenbaum, 1996).
- **Example:** If a teledermatology system starts with a limited dataset and is deployed for a specific demographic group, users from other demographics may cause the system to make inaccurate or biased decisions, as the system was not trained to account for their skin types or conditions.
- **PASSION Relevance:** Emergent bias could be a significant concern in PASSION, especially as the system expands to a larger, more diverse user base. If the platform's initial training data predominantly comes from one demographic, the system may perform less effectively for other skin types or conditions, leading to biased diagnosis or treatment recommendations.
- **Mitigation Strategy:** Continuous monitoring of how the system interacts with different demographic groups is essential. Ensuring that new data from diverse populations is incorporated into the training set periodically can help counteract emergent biases.

B.10 Category: External Influence Biases

External influence biases are introduced by external factors such as inappropriate benchmarks, reference tests, or popularity metrics. These factors can distort model predictions or evaluations, leading to biases in the system's decision-making process.

B.10.1 Evaluation Bias, *medium*

- **Definition:** Evaluation bias occurs when inappropriate or disproportionate benchmarks are used to assess the performance of a model. This can introduce external biases into the system by measuring it against benchmarks that don't fully represent the target data or user population (Suresh & Guttag, 2021; Buolamwini & Gebru, 2018).
- **Example:** If PASSION's dermatological model is evaluated using a benchmark set that overrepresents certain types of skin diseases, it may lead to the underperformance of the model for conditions that are less frequently represented in the benchmark.
- **PASSION Relevance:** This bias is relevant to PASSION because the dermatological datasets used to train and evaluate the model must be diverse and representative of the broader population. An evaluation benchmark skewed toward common conditions could impair the model's ability to accurately diagnose rare or underrepresented skin diseases.
- **Mitigation Strategy:** PASSION should implement diverse and representative benchmarks to evaluate model performance, ensuring that rare or less common conditions are also included in the evaluation dataset. Regular updates to the evaluation set as the dataset grows will help mitigate evaluation bias.

B.10.2 Incorporation Bias, *low*

- **Definition:** Incorporation bias arises when index tests in diagnostic accuracy studies are part of the reference tests, leading to artificially elevated sensitivity for the index tests (**c21**; **c25**; **c26**; Chakraborty, 2024; Young et al., 2020).
- **Example:** If PASSION uses diagnostic tests that are part of its reference set for evaluating accuracy, this could result in an overestimation of the model's sensitivity because the model is essentially being compared to itself, skewing results.
- **PASSION Relevance:** Incorporation bias is less relevant for PASSION since the platform likely relies on independent diagnostic benchmarks and tests to validate its dermatological models, reducing the chance of this type of bias affecting its evaluations.
- **Mitigation Strategy:** Ensuring that the reference tests used for validation are distinct and independent from the model's diagnostic tests can mitigate

incorporation bias.

B.10.3 Popularity Bias, *low*

- **Definition:** Popularity bias occurs when more popular items or data points are exposed more often in the training dataset or evaluation process. This can lead to a model that overemphasizes popular features or outcomes, disregarding less common but potentially important cases (Ciampaglia et al., 2018; Mehrabi et al., 2021).
- **Example:** In the context of PASSION, if the training data is overly focused on commonly encountered dermatological conditions or frequently observed features, the model may struggle to correctly diagnose rarer skin diseases that are underrepresented.
- **PASSION Relevance:** Popularity bias is relevant for PASSION, particularly if the training dataset includes a disproportionate number of common skin conditions, thereby reducing the effectiveness of the model for rarer conditions.
- **Mitigation Strategy:** To mitigate popularity bias, it is important for PASSION to ensure that the training dataset includes a balance of both common and rare skin conditions, offering a comprehensive representation of dermatological diseases.

B.11 Category: Cognitive Biases

Cognitive biases refer to systematic patterns of deviation from norm or rationality in judgment, whereby inferences about other people and situations may be drawn in an illogical fashion. These biases can impact how data is presented and interpreted (Mester, 2017).

B.11.1 Confirmation Bias, *high*

- **Definition:** Confirmation bias occurs when individuals favor information that confirms their preconceptions, leading them to ignore or dismiss evidence that contradicts their beliefs (Mester, 2017).
- **Example:** In healthcare, patients may interpret their symptoms based on information they find on the internet, confirming their own beliefs about a condition, even if this information is not medically accurate (**c15**; **c14**; Chakraborty, 2024).
- **PASSION Relevance:** For PASSION, confirmation bias could affect the initial diagnoses of dermatological conditions, resulting in biased labeling of skin diseases. If a medical professional has preconceived notions about a condition, they may incorrectly diagnose or label skin diseases, influencing the quality and accuracy of data.

- **Mitigation Strategy:** To reduce confirmation bias, diagnostic labels in PASSION could be cross-checked by multiple independent experts, ensuring diverse viewpoints and reducing the impact of pre-existing biases on data labeling.

B.11.2 Belief Bias, *high*

- **Definition:** Belief bias occurs when an individual's judgment is unduly influenced by their pre-existing beliefs or intuitions, leading them to accept conclusions that fit those beliefs without critically evaluating the evidence (Mester, 2017).
- **Example:** A researcher may ignore contradictory data in favor of results that support their hypothesis, even when the data doesn't robustly support their claim (Mester, 2017).
- **PASSION Relevance:** In the context of PASSION, belief bias could lead to inaccurate diagnosis and labeling if experts rely too heavily on their subjective interpretation of the data rather than objectively evaluating it. This could skew the dataset, impacting model training and accuracy.
- **Mitigation Strategy:** Implementing blind labeling processes, where experts are unaware of previous diagnoses, could help reduce belief bias. Additionally, training experts to focus on evidence-based diagnostic criteria would help mitigate the impact of this bias.

B.11.3 Previous Opinion Bias, *medium*

- **Definition:** Previous opinion bias occurs when the knowledge of prior results or diagnoses influences the interpretation of new data, leading to biased conclusions (Chakraborty, 2024).
- **Example:** A dermatology expert who knows the result of a previous diagnosis might let this knowledge influence their interpretation of subsequent test results, leading to potential bias in the diagnosis process (Chakraborty, 2024).
- **PASSION Relevance:** In PASSION, this bias could affect the consistency and accuracy of dermatological diagnoses. If experts are aware of previous diagnoses, they might be influenced by them, which could compromise the reliability of data in the system.
- **Mitigation Strategy:** To reduce this bias, PASSION could ensure that labeling experts independently diagnose cases without access to previous diagnoses, promoting impartiality in each evaluation.

B.11.4 Cause-Effect Bias, *low*

- **Definition:** Cause-effect bias arises when correlations between two variables are incorrectly interpreted as indicating a causal relationship, even when no such relationship exists (Mester, 2017).

- **Example:** An increase in the occurrence of skin rashes may be correlated with a particular season, but mistakenly concluding that the season is the cause of the rashes, rather than other factors, would be an example of cause-effect bias (Mester, 2017).
- **PASSION Relevance:** Cause-effect bias is less of an issue in PASSION, since the dataset primarily deals with diagnoses and symptoms without analyzing the underlying causes of diseases. However, if the algorithm were to be trained to predict causes, there could be a risk of misinterpreting correlations as causal relationships.
- **Mitigation Strategy:** To prevent cause-effect bias, any future development in PASSION's algorithm should focus on clear differentiations between correlation and causation, ensuring that predictions are based on robust, validated data.

B.11.5 Historical Bias, *high*

- **Definition:** Historical bias refers to biases that exist in the world or society, which can influence data collection and generation processes. These biases are often a reflection of past societal inequities (Suresh & Guttag, 2021).
- **Example:** A dataset that primarily includes images of skin conditions from a specific demographic (e.g., primarily white individuals) may not accurately represent skin diseases in other populations (Mehrabi et al., 2021).
- **PASSION Relevance:** Historical biases in the dermatology field, such as underrepresentation of certain skin types in clinical studies, could affect the quality of the PASSION dataset. This could lead to algorithms that perform poorly for underrepresented groups.
- **Mitigation Strategy:** Ensuring diversity in the dataset by collecting data from a wide range of demographic groups (age, gender, race, etc.) is essential to reduce historical bias in PASSION's dataset. Efforts should be made to balance the dataset and account for historically marginalized groups.

B.11.6 Content Production Bias, *medium*

- **Definition:** Content production bias occurs when biases are introduced during the creation of user-generated content, influenced by the creators' backgrounds, contexts, or perspectives (Olteanu et al., 2019).
- **Example:** In a study, images of skin diseases may be taken by healthcare professionals in settings that differ from those where the disease is most prevalent, leading to a potential misrepresentation of the condition's typical appearance (Olteanu et al., 2019).
- **PASSION Relevance:** In the context of PASSION, content production bias could arise in how images of skin diseases are taken. Variations in lighting, angle, or the quality of images could lead to inconsistencies, which may affect the training and performance of machine learning models.

- **Mitigation Strategy:** To reduce content production bias, standardization of image collection protocols could be implemented, ensuring consistent lighting, angles, and image quality. Additionally, training experts to adhere to these standards would help minimize bias in the data collection process.

B.12 Category: Behavioral Biases

Behavioral biases occur due to the actions and judgments of individuals, which are influenced by cultural, contextual, and platform-related factors. These biases can affect data collection, interpretation, and conclusions (Olteanu et al., 2019).

B.12.1 Behavioral Bias, *medium*

- **Definition:** Behavioral bias refers to how individuals' behavior can be influenced by the platforms they interact with, their cultural background, or their personal context (Olteanu et al., 2019).
- **Example:** Patients from different countries may present different behaviors when seeking medical advice for skin conditions, influenced by their cultural background and understanding of healthcare (Olteanu et al., 2019).
- **PASSION Relevance:** For PASSION, behavioral biases could influence who seeks dermatological care and why. Differences in healthcare-seeking behavior across cultures or countries may lead to an unrepresentative sample in the dataset. Therefore, including data from various countries could help account for these differences and improve the generalizability of the model.
- **Mitigation Strategy:** To mitigate behavioral bias, PASSION should aim to include a diverse set of data points from various geographical and cultural backgrounds. This would help ensure that the model is representative of different healthcare-seeking behaviors.

B.12.2 Self-Selection Bias, *high*

- **Definition:** Self-selection bias occurs when participants in a study are allowed to choose whether to participate, leading to an unrepresentative sample where certain groups are over- or underrepresented (Mester, 2022; Mehrabi et al., 2021).
- **Example:** In PASSION, only patients who seek dermatological care at hospitals would be included, which could exclude individuals with skin conditions who do not seek medical help, leading to skewed data (Mester, 2022; Mehrabi et al., 2021).
- **PASSION Relevance:** Self-selection bias is a significant issue for PASSION since the dataset relies on patients who visit hospitals, meaning those who do not seek treatment or who do not have access to healthcare will be underrepresented in the dataset.

- **Mitigation Strategy:** To mitigate self-selection bias, PASSION could look into alternative data sources, such as surveys or community outreach programs, to gather information from individuals who may not seek formal dermatological care.

B.13 Category: Publication Biases

Publication biases are introduced when research outcomes are selectively reported or published based on certain characteristics such as positive results or trending topics. These biases can distort the scientific record and lead to misinterpretation or overemphasis on particular findings.

B.13.1 Publication Bias, *high*

- **Definition:** Publication bias occurs when studies with significant or positive results are more likely to be published than studies with non-significant or negative results. This leads to a skewed representation of the effectiveness of an intervention or treatment.
- **Example:** If studies showing positive results of a dermatological treatment for skin diseases are more likely to be published than studies with neutral or negative findings, this creates a publication bias in the medical literature.
- **PASSION Relevance:** In the context of the PASSION dermatology dataset, publication bias may manifest if studies based on the dataset predominantly focus on successful diagnoses or treatments, leaving out less effective or inconclusive results. This can lead to an overestimation of the dataset's utility and effectiveness in detecting skin diseases.
- **Mitigation Strategy:** A strategy to mitigate publication bias is the promotion of open access to all research outcomes, including negative or neutral results. Encouraging the publication of replication studies and meta-analyses that incorporate a wide range of findings, not just the most positive ones, can help counteract this bias.

B.13.2 Hot Stuff Bias, *medium*

- **Definition:** Hot stuff bias refers to the tendency for journals to be less critical of research related to trending or highly popular topics, leading to the disproportionate publication of these studies.
- **Example:** In dermatology, if there is a sudden interest in a new skin disease detection technology, studies related to this technology may be published more frequently, regardless of their quality, because they align with the current hot topic.
- **PASSION Relevance:** In the context of PASSION, if a certain skin disease detection method is trending, there could be a tendency for studies utilizing

the PASSION dataset to be published more often, potentially overshadowing other important findings or datasets that may also contribute valuable insights.

- **Mitigation Strategy:** To mitigate hot stuff bias, it is important to prioritize the quality and robustness of the research rather than focusing on its alignment with current trends. Peer reviewers should be vigilant and ensure that the novelty of a topic does not overshadow the scientific rigor of the study.

B.13.3 All is Well Bias, *low*

- **Definition:** All is well bias occurs when theories that align with the majority or dominant views are more likely to be published than those that challenge the consensus.
- **Example:** In the field of dermatology, if the majority of researchers agree on a specific method for diagnosing skin diseases, studies questioning the effectiveness of this method may be less likely to be published, even if they provide valid and critical insights.
- **PASSION Relevance:** This bias is less directly relevant to PASSION as the dataset itself is focused on real-world data collection, which may be less influenced by theoretical debates. However, if there is widespread consensus on a particular model or diagnostic approach using PASSION data, studies presenting opposing findings could be underrepresented.
- **Mitigation Strategy:** Encouraging diversity in research perspectives and methodologies can help counteract the all is well bias. Peer reviewers should actively look for and support research that challenges the majority view, ensuring that minority perspectives are also given a platform.

B.13.4 Rhetoric Bias, *medium*

- **Definition:** Rhetoric bias occurs when the way in which research findings are presented, such as through charismatic writing or media coverage, influences the perception of those findings more than the actual data.
- **Example:** If a particular dermatology treatment is presented with highly persuasive language and strong media support, it may be perceived as more effective than it truly is, regardless of the underlying data.
- **PASSION Relevance:** In the context of PASSION, if studies using the dataset are written in a particularly persuasive or charismatic manner, they may attract more attention, regardless of their scientific rigor. This could lead to an overemphasis on certain findings, overshadowing others that may be equally valuable but presented less dramatically.
- **Mitigation Strategy:** Researchers should focus on presenting data clearly and objectively, avoiding exaggerated claims or overly persuasive language. Journals and reviewers should also ensure that rhetoric does not overshadow the actual scientific contribution of a paper.

B.13.5 Novelty Bias, *high*

- **Definition:** Novelty bias refers to the tendency to favor new interventions, treatments, or findings over established ones, often because newer approaches are perceived as being better, even if the evidence does not support this.
- **Example:** A new method for detecting skin diseases, such as a machine learning model, may be hailed as a breakthrough, even if it has not been proven to be more effective than traditional diagnostic methods.
- **PASSION Relevance:** Novelty bias can be particularly relevant for PASSION as researchers may place disproportionate emphasis on the latest machine learning techniques or models, potentially overlooking more established methods that are still effective in skin disease detection. This bias can lead to a focus on novelty at the cost of reliability.
- **Mitigation Strategy:** To mitigate novelty bias, researchers should compare new approaches to established methods in rigorous, controlled studies. Peer reviewers should ensure that novelty does not overshadow the importance of replicability and robustness in research findings.

Potential Biases in PASSION These biases are relevant for all researchers working with datasets like PASSION. They should be kept in mind when interpreting, publishing, and peer-reviewing papers. Ensuring that these biases are recognized and addressed will help maintain the integrity and utility of the PASSION dermatology dataset in advancing skin disease detection and treatment.

B.14 Category: Medical Biases

Medical biases refer to distortions in healthcare and diagnostic practices that can arise due to various external factors, leading to an inaccurate representation of diseases or patient conditions. These biases can affect both clinical assessments and datasets, potentially skewing the analysis and treatment approaches.

B.14.1 Popularity Bias, *high*

- **Definition:** Popularity bias occurs when more well-known or stigmatized diseases are over-represented in healthcare settings compared to less common diseases. This can result in a distorted view of the prevalence and severity of different conditions.
- **Example:** A hospital may see more cases of diseases that are widely known or stigmatized, such as skin cancers, leading to an overemphasis on these conditions in research and treatment, while conditions like rare dermatological disorders are underrepresented.
- **PASSION Relevance:** Popularity bias is highly relevant to the PASSION dataset, as it may include more common or well-known skin diseases due to hospital reporting trends. This could lead to an over-representation of these

conditions in the dataset, skewing the model's ability to detect less prevalent dermatological diseases accurately.

- **Mitigation Strategy:** To mitigate popularity bias, PASSION should ensure a balanced representation of diseases in the dataset. This could involve actively seeking data from hospitals or clinics that treat a wider variety of dermatological conditions, including rare ones.

B.14.2 Apprehension Bias, *low*

- **Definition:** Apprehension bias arises when patients exhibit anxiety or fear about upcoming medical procedures, which can influence physiological measurements or diagnostic results, leading to inaccuracies.
- **Example:** A patient may have elevated blood pressure readings due to anxiety before a dermatological procedure, leading to an inaccurate diagnosis or assessment.
- **PASSION Relevance:** While apprehension bias may be relevant in clinical settings, its direct impact on the PASSION dataset is likely low. The dataset focuses on dermatological conditions, where apprehension bias is less likely to affect diagnostic images or disease annotations.
- **Mitigation Strategy:** Although not a primary concern for PASSION, ensuring that patients are comfortable during the data collection process and minimizing procedural anxiety could improve the accuracy of any associated clinical measurements.

B.14.3 Hawthorne Bias, *medium*

- **Definition:** Hawthorne bias refers to changes in behavior by subjects when they know they are being observed, which can influence study results and clinical assessments.
- **Example:** If clinicians or patients are aware that their cases are being monitored for a dermatology study, they might alter their behavior, such as reporting symptoms differently or providing more detailed information than they normally would.
- **PASSION Relevance:** The Hawthorne bias could be relevant in the PASSION dataset, particularly if annotators or clinicians are aware that their diagnostic decisions are being evaluated. This could lead to over-reporting or under-reporting certain disease characteristics.
- **Mitigation Strategy:** PASSION could utilize regular follow-ups to observe natural trends in disease progression and symptom reporting. By minimizing the knowledge of when they are being observed, PASSION can reduce the impact of this bias on the data quality.

B.14.4 Centripetal Bias, *medium*

- **Definition:** Centripetal bias occurs when patients tend to seek care from well-known or highly reputable specialists or institutions, which may skew the cases seen by those professionals towards more complex or specialized conditions.
- **Example:** In dermatology, patients with severe or uncommon conditions may prefer to see a renowned specialist, while more routine cases are handled by general practitioners or less well-known dermatologists.
- **PASSION Relevance:** Centripetal bias is relevant to PASSION as it may affect which hospitals or specialists contribute data to the dataset. If more well-known institutions are over-represented, this could lead to a dataset that is biased towards more severe or complicated cases of skin diseases.
- **Mitigation Strategy:** PASSION can mitigate centripetal bias by ensuring diversity in the data sources. It should include data from both specialized and general dermatology practices, as well as a mix of both urban and rural hospitals, to provide a comprehensive view of skin diseases across different settings.

Potential Biases in PASSION PASSION must be careful in interpreting the metadata. Since the data is from hospitals, there could be an over-representation of more popular or severe diseases, leading to a distorted dataset. Additionally, PASSION could leverage the Hawthorne bias to improve the consistency and quality of the annotations, using follow-ups to observe how annotators' behavior may shift over time. Furthermore, centripetal bias can be considered when selecting partners to work with, ensuring that a variety of specialists and institutions contribute to the dataset.

B.15 initial sources

TODO: check that all stuff above matches the stuff below

B.15.1 Bias Introduction

- **Assessment Tools** An interesting direction that researchers have taken is introducing tools that can assess the amount of fairness in a tool or system. For example, Aequitas [136] is a toolkit that lets users to test models with regards to several bias and fairness metrics for different population subgroups. Aequitas produces reports from the obtained data that helps data scientists, machine learning researchers, and policymakers to make conscious decisions and avoid harm and damage toward certain populations. AI Fairness 360 (AIF360) is another toolkit developed by IBM in order to help moving fairness research algorithms into an industrial setting and to create a benchmark for fairness algorithms to get evaluated and an environment for fairness researchers to share their ideas [11]. These types of toolkits can be helpful for learners, researchers, and people working in the industry to move towards developing fair machine learning application away from discriminatory behavior (Mehrabi et al., 2021).

- Most AI systems and algorithms are data driven and require data upon which to be trained. Thus, data is tightly coupled to the functionality of these algorithms and systems. In the cases where the underlying training data contains biases, the algorithms trained on them will learn these biases and reflect them into their predictions. As a result, existing biases in data can affect the algorithms using the data, producing biased outcomes. Algorithms can even amplify and perpetuate existing biases in the data.(Mehrabi et al., 2021).
- In addition, algorithms themselves can display biased behavior due to certain design choices, even if the data itself is not biased. The outcomes of these biased algorithms can then be fed into real-world systems and affect users' decisions, which will result in more biased data for training future algorithms.(Mehrabi et al., 2021).
- Bias can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks. In (Suresh & Guttag, 2021), authors talk about sources of bias in machine learning with their categorizations and descriptions in order to motivate future solutions to each of the sources of bias introduced in the paper. In (Olteanu et al., 2019), the authors prepare a complete list of different types of biases with their corresponding definitions that exist in different cycles from data origins to its collection and its processing.(Mehrabi et al., 2021).
- The list of biases that can occur in any research is considerably long, and certainly not all of them can be avoided. (Chakraborty, 2024)

B.15.2 Biases Extensive Sources

Data Biases

Data biases (data to algorithm (biases in data which might have an impact on biased algorithmic outcomes (Mehrabi et al., 2021)))

-

Algorithmic Biases

Algorithmic biases (Algorithm to user (A modulates U behaviour, biases in algorithm might lead to introduce biases in user behaviour and affect it as a consequence)) (Mehrabi et al., 2021)

User Biases

User to Data (user-generated data, inherent biases in users could be reflected in the data they generate; biases in last section might introduce further bias in this process) (Mehrabi et al., 2021)

Dermatology Biases

- Equity. AI has the potential to worsen health-care disparities, as recognized by the popular media (Khullar, 2019), particularly in dermatology (Adamson and Smith, 2018). The first concern is adequate representation of underserved populations in training data. Existing DL models have been trained on mainly European or East Asian populations, and the relative lack of training on darker skin pigmentation may limit overall diagnostic accuracy. This possibility is demonstrated by the increased error rates in commercial systems, trained on predominantly white datasets, for facial analysis in identifying black individuals (Buolamwini and Gebru, 2018). Second, AI may entrench existing social and economic biases and perpetuate inadvertent discriminatory practices, for example, in recommending less follow-up for black patients than for whites, when health costs are used as a proxy for health needs (Obermeyer et al., 2019). Third, disproportionate adoption by different groups may exacerbate existing inequities. Access to and use of technology differs based on sociodemographics (Tsetsi and Rains, 2017), and more techsavvy users may be more likely to embrace AI for skin screening (Tong and Sopory, 2019). The issue of equity in AI diagnosis needs to be carefully addressed to avoid inadvertent exacerbation of health-care disparities. (Young et al., 2020) - dermatology
- Model generalizability. Generalizability is a major concern for AI models; studies of computer-assisted diagnosis of melanoma report lower sensitivity for melanoma on independent test sets than on nonindependent test sets (Dick et al., 2019). It is difficult to study generalizability because published DL models are not publicly available, making it impossible to compare performance, unless each study uses a standardized benchmark database, such as the Melanoma Classification Benchmark (Brinker et al., 2019d). Han et al. (2018a) reported excellent metrics of performance and made their model available for image submission; however, the model prediction was not robust when images from an outside clinic were submitted, image magnification or contrast was altered, or images were rotated (NavarreteDechent et al., 2018). On ImageNet, a nonmedical dataset of 1,000 object categories, training on a dataset of 300 ... (Young et al., 2020)

B.15.3 Bias Sources

The general ML lifecycle consists of data gathering, training the algorithm and the user interaction with the trained model. Now, while data gathering, biases can arise either through the collection process or it is already inherited in the available data. Further, depending on the algorithm design, during training, the existing bias in the data can be amplified and new bias can be introduced. Lastly, the result of the algorithm can affect the user experience on inference which can lead to further bias amplification. This generates a feedback loop between the biases in each step of the ML lifecycle which can make it hard to identify the original bias source. The feedback loop is illustrated in Figure B.1, which also shows first bias definitons, which were categorized according to this feedback loop (Mehrabi et al., 2021).

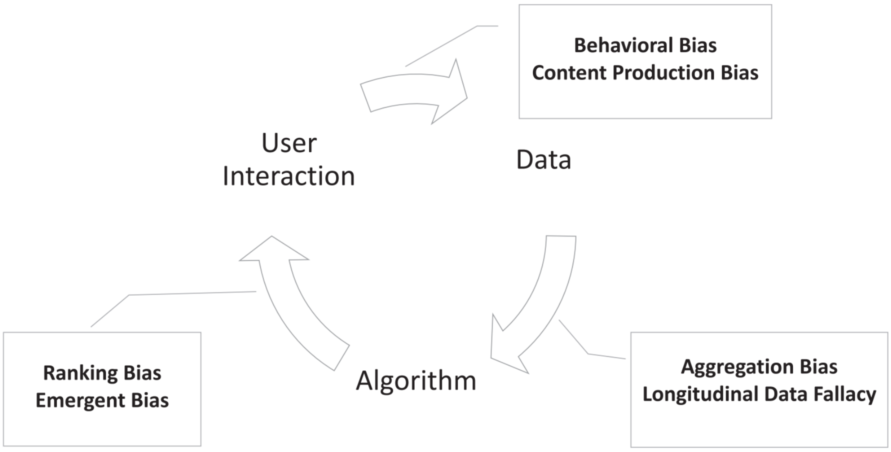


Figure B.1: Bias definitions in a ML lifecycle (Mehrabi et al., 2021).

- two potential sources of unfairness in machine learning outcomes - those that arise from biases in the data and those that arise from the algorithms ... we observe that biased algorithmic outcomes might impact user experience, thus generating a feedback loop between data, algorithms and users that can perpetuate and even amplify existing sources of bias (Mehrabi et al., 2021).
- The loop capturing this feedback between biases in data, algorithms, and user interaction is illustrated in Figure 1. We use this loop to categorize definitions of bias in the section below (Mehrabi et al., 2021)

B.15.4 Bias Types

The Table B.1 aims to provide an overview over what kind of biases exist according to research. The more detailed categories listed in the table try to capture similar kind of biases. This thesis follows roughly the categorization of Mehrabi et al. (2021). Some biases might acctually fit in multiple categories. The definition of the categories including examples of specific biases follows.

TODO: check the c citations in the following chapters

Bias	**Mentioned in Context of**	
	ML	**Dermatology**
Data Biases		
Sampling Biases	X[1,2,3]	X[4]
Representation Biases	X[1]	X[5,6]
Measurement Biases	X[1,3]	X[4,6]
Research Biases	X[7]	X[4]
Feature Representation Biases	X[1,3]	X[4]
Imaging Biases		X[5]
Medical Biases	X[8]	X[4]
Temporal Data Biases	X[1]	X[4]
Algorithmic Biases		
User-Algorithm Interaction Biases	X[1]	
External Influence Biases	X[1]	X[4]
User Biases		
Cognitive Biases	X[1,7]	X[4]
Behavioral Biases	X[1,3]	X[4,5]
Publication Biases		X[4]
Medical Biases	X	X[4]

[1] (Mehrabi et al., 2021)
[2] (HP, 2022)
[3] (Mester, 2022)
[4] (Chakraborty, 2024)
[5] (Young et al., 2020)
[6] (Montoya et al., 2025)
[7] (Mester, 2017)
[8] (Delgado-Rodríguez & Llorca, 2004)

Table B.1: Bias categories - grouped according the ML lifecycle of Mehrabi et al. (2021)

TODO: Feedback Astrid: zu viele Unterkapitel -> anders strukturieren

B.15.4.1 Data Biases

Sampling Biases

When gathering data, it's usually not possible to gather the data of a whole population. Instead, the data is gathered by sampling. A sample is a subgroup of individuals from the population. To get unbiased results, this sampling process should represent the true population, with a low sampling error (HP, 2022). This is often achieved with randomized samples. With non-random sampling processes, sampling bias arises. The consequence is, that the insights of one sampled population may not generalize with insights on another sampled popluation (Mehrabi et al., 2021).

Those biases can be introduced with a flawed sampling process:

- **Sampling bias**, due to nonrandom sampling of subgroups, leading to poor generalization (Mehrabi et al., 2021)
- **Selection bias**, working only on specific subset of the population which is not representative (Mester, 2022; Chakraborty, 2024)

- **Systematic selection bias**, chosen samples differ dramatically from the representative populations; e.g. in dermatology, when only the most severe patient data gets included (**c5**; **c6**; **c33**; Chakraborty, 2024)
- **Ascertainment bias**, tendency to exclude segments from the population due to e.g. cultural differences, such as which patient segment goes to government clinics vs. private clinics (usually influenced by socioeconomic status) (**c5**; Chakraborty, 2024)
- **Availability bias**, focus on widely available data instead of most representative data (**c9**; **c10**; Chakraborty, 2024**<empty citation>**)
- **Survivorship bias**, focus only on pre-selected data, ignoring the initial data-points which got filtered out (Mester, 2022).

Potential Biases in PASSION PASSION tries to reduce sampling bias in dermatology against high pigmented skin. PASSION might introduce (systematic) selection bias or Ascertainment bias, if in the dermatology centers only sickest / more severe patients are seen as indicated by Chakraborty (2024) PASSION inherits availability bias as it is using FST scale. Survivorship bias could be relevant for PASSION, if dermatology diseases could be lethal. Further, all patients which are not able to go to one of the dermatology centers which were used in PASSION could be considered to left out by survivorship bias.

used

- Sampling Bias. Sampling bias is similar to representation bias, and it arises due to nonrandom sampling of subgroups. As a consequence of sampling bias, the trends estimated for one population may not generalize to data collected from a new population. (Mehrabi et al., 2021). This is what the PASSION dataset tries to improve
- Selection bias - wrong sampling method, working on a specific subset of audience; usually by working only with data that is easy to access (Mester, 2022; Mester, 2017) - statistical bias
- Selection bias: Since it is not possible to work with large populations, for most dermatological studies, samples are chosen that are said to be representative of the original population. In selection bias, the selected subgroups are not representative of their original population. A variation of this is systematic selection bias, where samples chosen differ dramatically from their representative populations. Our experience suggests, such selection bias occurs more commonly in studies conducted in regional referral centers where only the sickest or more severe patients are usually seen. For example, a study compared the efficacy of thalidomide vs. prednisolone in hospitalised patients of erythema nodosum leprosum. It derived that thalidomide was more efficacious than steroids in erythema nodosum leprosum. Such findings cannot be generalised to all erythema nodosum leprosum since patients admitted to a regional referral center will likely have more severe disease.5,6,33 (Chakraborty, 2024)
- Availability bias: More emphasis is placed on widely available data than

scantily available data. A classic example is the use of antihistamines in pregnancy dermatoses, where nearly all standard books recommend first-generation antihistamine chlorpheniramine because more data is available.9 10. (Chakraborty, 2024) - dermatology

- Survivorship bias (Mester, 2022; Mester, 2017) - statistical bias
- Ascertainment Bias: This bias is commonly encountered in venereology practice. It is defined as a bias due to the tendency of some segments of the target population to get excluded due to cultural and other differences. For example, in most venereology clinics in government setups, studies show that venereal diseases are commoner in lower socioeconomic status. One reason might be that the higher socioeconomic status people tend to go to private practitioners and thereby get excluded from government-run clinics.9,10 Allocation concealment and blinding are good ways to avoid this. 5. (Chakraborty, 2024) - healthcare

even more extensive

- Selection bias is again divided into two types endogenous selection bias and exogenous selection bias. The best example of endogenous selection bias in dermatology is the inclusion of non-response. If a trial tests the efficacy of a particular biologic in psoriasis, the response is usually collected from trial participants via postal services. Certain participants will not respond, although they might have substantially improved. Their exclusion will result in significant differences in efficacy evaluation.33 Exogenous selection bias results when both treatment and outcome result from dependency on an external variable that is not controlled. For example, if sunlight exposure is not controlled, it will influence both the intervention and control groups since psoriasis is a photosensitive (and photoexcerbated) dermatosis. (Chakraborty, 2024) - dermatology
- survivorship bias - World War II planes (**Silfwer_2017**) - https://doctorspin.org/media-psychology/psychology/survivorship-bias/

Representation Biases

TODO: still describe this category

Those biases can be introduced :

- **Representation bias**, non-representative sample lead to missing subgroups or other representation anomalies, which can be harmful to downstream applications. Popular ML datasets suffer from representation bias (Mehrabi et al., 2021; Shankar et al., 2017)
- **Population Bias**. Population bias arise when statistics, demographics and characteristics in the sample differ from the target population (Olteanu et al., 2019). The data it creates is non-representative for the target population (Mehrabi et al., 2021).

- **Aggregation bias** occurs, when ”false conclusions are drawn about individuals from observing the entire population”. It doesn’t matter, whether the subgroups are represented equally in the training set, any generalized assumptions can result in aggregation bias (Mehrabi et al., 2021). In medicine, diseases can present themselves differently across genders and ethnicities (Suresh & Guttag, 2021). Therefore, diagnostic models need to incorporate those differences to mitigate aggregation bias (Mehrabi et al., 2021).
- **Simpson’s Paradox** is a type of aggregation bias, which arises in heterogeneous data analysis. Observed associations disappear or reverses in the subgroup data (Mehrabi et al., 2021).

Potential Biases in PASSION PASSION tries to mitigate representation bias, by including more FST skin types - however, it could introduce other representation biases Aggregation bias and Simpson’s Paradox could potentially be an issue when the analyzed skin diseases present themselves differently in patients based on their genetics

used

- Representation Bias. Representation bias arises from how we sample from a population during data collection process (Suresh & Guttag, 2021). Non-representative samples lack the diversity of the population, with missing subgroups and other anomalies (Mehrabi et al., 2021).
- Popular machine-learning datasets that serve as a base for most of the developed algorithms and tools can also be biased—which can be harmful to the downstream applications that are based on these datasets. ... In (Shankar et al., 2017), researchers showed that these datasets suffer from representation bias and advocate for the need to incorporate geographic diversity and inclusion while creating such datasets. (Mehrabi et al., 2021)
- Population Bias. Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population (Olteanu et al., 2019). Population bias creates non-representative data. ... More such examples and statistics related to social media use among young adults according to gender, race, ethnicity, and parental educational background can be found in (Hargittai, 2007). (Mehrabi et al., 2021)
- Aggregation Bias. Aggregation bias (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population. An example of this type of bias can be seen in clinical aid tools. Consider diabetes patients who have apparent morbidity differences across ethnicities and genders. Specifically, HbA1c levels, that are widely used to diagnose and monitor diabetes, differ in complex ways across genders and ethnicities. Therefore, a model that ignores individual differences will likely not be well-suited for all ethnic and gender groups in the population (Suresh & Guttag, 2021). This is true even when they are represented equally in the training data. Any general assumptions about subgroups within the population-

lation can result in aggregation bias. (Mehrabi et al., 2021). –> could also be important for dermatology issues!!!

  - Simpson's Paradox. Simpson's paradox is a type of aggregation bias that arises in the analysis of heterogeneous data [18]. The paradox arises when an association observed in aggregated data disappears or reverses when the same data is disaggregated into its underlying subgroups (Fig. 2(a)). ... After analyzing graduate school admissions data, it seemed like there was bias toward women, a smaller fraction of whom were being admitted to graduate programs compared to their male counterparts. However, when admissions data was separated and analyzed over the departments, women applicants had equality and in some cases even a small advantage over men. The paradox happened as women tended to apply to departments with lower admission rates for both genders. Simpson's paradox has been observed in a variety of domains, including biology [37], psychology [81], astronomy [109], and computational social science [91].(Mehrabi et al., 2021).

Measurement Biases

How features are chosen, used and measured can lead to biases (Mehrabi et al., 2021; Suresh & Guttag, 2021).

Examples for such biases are:

- **Measurement bias** in general, e.g. using mismeasured proxy variables lead to misinterpretations of the outcome (Mehrabi et al., 2021)
- **Observer bias** is a subconscious bias which can occur in different forms. Either, researchers projects their own expectations on the research and influence the testers accordingly (Mester, 2022). In other cases, different observes report the same observation differently (**c29**; **c26**; Chakraborty, 2024)
- **Annotator bias** is a special form of observer bias. The labeling process of human annotators can be influenced by lots of factors (e.g. personal background, social context) and even minor design choices (e.g. scale order, image context). This can introduce inconsistencies when labeling the data (Montoya et al., 2025)
- **Recall bias**. This bias occurs when queried individuals do not remember things correctly, due to humans selective memory. This can cause misinterpretation, for example when analyzing causes and effects of behaviour on certain diseases in medicine (Mester, 2022**c3-6**; **c2**; Chakraborty, 2024).

Potential Biases in PASSION Measurement Bias (proxy var) - Country of Origin in PASSION depending on the interpretation - should not be used for ethnicity, as this is not linked directly to the genes, see example https://medium.com/bcggamma/prac
ai-responsibly-with-proxy-variable-detection-42c2156ad986

Annotator bias regarding skin tone labeling has been investigated in (Montoya et al., 2025). PASSION should evaluate its process.

used

- Measurement Bias. Measurement, or reporting, bias arises from how we choose, utilize, and measure particular features (Suresh & Guttag, 2021) (e.g. mismeasured proxy variables) (Mehrabi et al., 2021). (= e.g. someone who lives at that postal code probably has this ethnicity); –> could that be an issue with the country of origin feature?
- This study found that while using skin tone instead of race for fairness evaluations in computer vision seems objective, the annotation process remains biased by human annotators. Untested scales, unclear procedures, and a lack of awareness about annotator backgrounds and social context significantly influence skin tone labeling. This study exposes how even minor design choices in the annotation process, like scale order (dark to light instead of light to dark) or image context (face or no face, skin lesion presence), can sway agreement and introduce uncertainty in skin tone assessments. ... The researchers emphasize the need for greater transparency, standardized procedures, and careful consideration of annotator biases to mitigate these challenges and ensure fairer and more robust evaluations in computer vision. (Montoya et al., 2025) - demographic dermatology bias
- Observer bias - projecting expectations onto the research (Mester, 2022; Mester, 2017) - statistical bias
- Observer bias: When different observers view the same observation, they report it differently e.g., different observers may give differing descriptions about subtle features in the histopathology report of a skin biopsy.29 26. (Chakraborty, 2024) - dermatology
- Recall bias - respondent doesn't remember things correctly; Recall bias is another common error of interview/survey situations. It happens when the respondent doesn't remember things correctly. It's not about bad or good memory – humans have selective memory by default. After a few years (or even a few days), certain things stay and others fade. It's normal, but it makes research much more difficult. TODO: keep an eye on this when recalling evidences!! (Mester, 2022; Mester, 2017) - statistical bias
- Memory or recall bias: This is a type of bias where sufferers of a disease, often termed cases, have a greater tendency to recall a particular habit than non-sufferers, viz controls. This results in an uneven distribution of risk factors between the cases and controls. An example of this would be a case-control study to evaluate the association between dental amalgam use and the development of oral lichen planus. Those with lichen planus are more likely to recall a history of dental amalgam use than those who do not have the disease. This difference in recall between a diseased cohort and control has resulted in difficulties in assessing the association between diet and many dermatological diseases – like milk and chocolate consumption and acne, fatty meals and psoriasis, sugary meals and psoriasis, agricultural exposure to insecticides and pemphigus and so on.3–6 2. (Chakraborty, 2024) - dermatology

Research Biases

TODO: consider to move at beginning / out of data biases Researchers and their processes can also be biased in multiple ways:

- **Funding / Sponsorship bias**, when a study is deliberately supporting those findings, which the sponsor expects (**c22**; Chakraborty, 2024; Mester, 2017)
- **Data dredging bias**. The statistical methods and model are chosen to provide a certain p-value, to improve the probability of the research hypothesis being true. TODO: consider to move this to an own reporting section (Chakraborty, 2024)
- **Hypothetical bias**. Hypothetical questions lead to responses that do not reflect, what interviewees would do in real life. (**c31**; **c28**; Chakraborty, 2024) TODO: isn't this a user bias instead?

Potential Biases in PASSION Since the PASSION dataset is already published, the research biases might already be introduced. It is not feasible during the duration of this thesis to make an evaluation on those biases. Instead, I would recommend the PASSION team and researcher in general, to check the list above carefully and take measures against them. Maybe, an external evaluation could help to detect and prevent those biases even better.

used

- Funding bias (Mester, 2022; Mester, 2017) - statistical bias
- Industry sponsorship bias: This has now been reclassified as conflict-of-interest bias. In short, the study deliberately supports the findings expected from it by its sponsors. 22.(Chakraborty, 2024) - dermatology
  Reporting biases
- Data dredging bias: It is an entirely avoidable bias. This is subdivided into two types – Fishing type and "P-value hacking" type. It involves using multiple statistical methods to get the desired p-value and selecting the statistical model that gives the p-value the author wants. This is "lamentably common" in dermatological research.16 To detect data dredging bias, always perform a "p-curve analysis" while performing a meta-analysis.17,18 Much emphasis is nowadays given to the confidence interval instead of the p-value, which gives an approximate idea of the range in which one can be 95% (or 90%, depending on the confidence interval chosen) sure that the result is correct. The confidence interval remains unaffected by p-value dredging. This subject has been reviewed in depth in recent works.18,19 15.(Chakraborty, 2024)
- Hypothetical bias: Many dermatological researches (and some life quality questionnaires like vitiQoL) use hypothetical questions – like "What would you do when some stranger asks you about your lesion?". The responses to these questions by the study participants often do not tally with what they would do in real life. This is called hypothetical bias and is avoided by adopting the ex-ante approach.31 28. (Chakraborty, 2024) - dermatology

Feature Representation Biases

Some of those biases are:

- **Omitted Variable Bias** arises when variables are not included in the model, which leads to situations for which the model is not ready for (Mehrabi et al., 2021; Mester, 2022)(Clarke, 2005; Riegg, 2008)(Mustard, 2003).
- **Collider Bias** Two variables can influence a common third variable, the collider variable. When sampling is restricted by this collider variable, it could lead to a distortion (**c4**; **c8**; **c9**; Chakraborty, 2024).

Potential Biases in PASSION The ethnicity is omitted in the PASSION dataset which could lead to issues See the medical section for more specific collider bias, maybe there could be others

used

- Omitted Variable Bias. Omitted variable bias4 occurs when one or more important variables are left out of the model (Clarke, 2005; Riegg, 2008)(Mustard, 2003). Something that the model was not ready for(Mehrabi et al., 2021). did not take into account (Mehrabi et al., 2021)
- Omitted variable bias (Mester, 2022; Mester, 2017) - statistical bias
- Collider Bias: This is an under-appreciated bias, and often confused with a confounder. This is especially seen in observational studies where it is defined as a distortion produced by the restriction of sampling by a collider variable. A collider variable is defined as one that has an independent effect on the outcome studied apart from the studied variable. In simpler terms, collider bias occurs when exposure and development influence a common third variable. That variable or collider is controlled by study design or in the analysis. An example is the observation that psoriasis patients tend to have more depression and anxiety disorders. Since severe psoriasis patients tend to get hospitalised and also get screened for mental health issues, a spurious association between them could have been obtained due to collider bias. The two variables viz psoriasis and depression converged, i.e., collided, into a single outcome – hospitalization.8,9 4. (Chakraborty, 2024) - dermatology

Imaging Biases

Dealing with images can lead to a whole other set of challenges, which can lead to biases. The challenges are for example technical variations in hardware and software but also differences in how images are gathered or what is in it (Young et al., 2020).

Those biases can be introduced :

- **Image Quality Bias**. The quality of an image (zoom level, focus, lightning) could be associated with the classification (Young et al., 2020)

- **Visual Artifact Bias**. Other artifacts, such as presence of hair or surgical ink markings on dermatology images, can decrease classification performance (**Winkler et al.**; **2019 & Bisla et al.**; **2019 (from Young_2020)**)
- **Field of View Bias**. What view is captured in the image can interfere with prediction quality what is it, consequence (**Mishra et al.**; **2019 from Young_2020**)

Potential Biases in PASSION The PASSION model could learn to associate unrelated visual effects, hair, body parts or image quality with a disease.

used

- Image quality. Several barriers to AI implementation in the clinic need to be overcome with regards to imaging (Figure 1). These include technical variations (e.g., camera hardware and software) and differences in image acquisition and quality (e.g., zoom level, focus, lighting, and presence of hair). For example, the presence of surgical ink markings is associated with decreased specificity (Winkler et al., 2019), field of view can significantly affect prediction quality (Mishra et al., 2019), and classification performance improves when hair and rulers are removed (Bisla et al., 2019). We have developed a method to measure how model predictions might be biased by the presence of a visual artifact (e.g., ink) and proposed methods to reduce such biases (Pfau et al., 2019). Poor quality images are often excluded from studies, but the problem of what makes an image adequate is not well studied. Ideally, models need to be able to express a level of confidence in a prediction as a function of image quality and appropriately direct a user to retake photos if needed. (Young et al., 2020) - dermatology

Medical Biases

In ML for health care, there are special medical versions of the mentioned biases as well as completely new biases. They require special attention, since they directly influence the diagnosis or treatment of a disease.

Those biases can be introduced:

- **Berkesonian bias** occurs in hospital-based studies when two variables influence hospital or clinical attendance independently. This can lead to a distorted estimation of the relationship between those variables because the study population of hospitalized patients is not representative of the whole population (**c3**; **c7**; Chakraborty, 2024)
- **Informed presence bias**, the probability to get screened for other diseases is higher for people who seek medical care. Like Berkesonian bias, this can lead to misleading interpretations of relationships between two diseases (**c27**; **c23**; Chakraborty, 2024)
- **Diagnostic access bias**, depending on the geographical location, individuals have better access to medical care. Therefore, their disease prevalence could appear to be higher and diseases could be diagnosed earlier. (**c19-c21**; Chakraborty, 2024)

- **Diagnostic reference test bias** is a **verification bias**, where not all individuals receive the same reference test for the diagnostic process, potentially leading to different diagnoses. (**c21**; Chakraborty, 2024)
- **Mimicry bias**, exposures to treatment options can cause a disease which presents itself similar to the study disease, which potentially creates misleading data (**c28**; **c25**; Chakraborty, 2024)
- **Unacceptable Disease bias**. When a disease is socially unacceptable, it can result in under-reporting of the same disease (**c30**; **c27**; Chakraborty, 2024)
- **Healthy volunteer selection bias**, is a type of self-selection bias where the volunteers are in general healthier than the population due to more interest in health (Delgado-Rodríguez & Llorca, 2004)

Potential Biases in PASSION Berkesonian bias depending on the chosen hospitals Informed presence bias regarding correlation between impedigo and the other diseases Diagnostic access bias can somewhat be addressed by PASSION, since its dataset includes samples of later states of diseases. However, in the PASSION context itself, this bias could still be relevant. Diagnostic reference test bias could be inherited in the PASSION dataset, depending on how the dermatologists work. Mimicry bias is not relevant regarding the exposures since PASSION does not hold any exposure data. However, diseases which mimicry others could lead to issues if they are not detected.

used

- Berkesonian Bias: Named after Dr. Joseph Berkeson, this bias reflects the variation in rates of hospital admission or clinic attendance for different diseases. For example, if a study is conducted to know the effect of pregnancy on syphilis in an antenatal clinic, we are likely to get biased data since the two conditions, viz pregnancy and syphilis, are both likely to affect clinic attendance and all observations related to the relationship between pregnancy and syphilis.7 3. (Chakraborty, 2024) - dermatology
- Informed presence bias: Simply, a person attending a health center is more likely to get screened for other unrelated comorbidities than those not attending a health center e.g., the finding psoriasis is associated with depression has now been criticised because those having psoriasis also have a greater chance to be screened for depression since they are already attending a health center.27 23. (Chakraborty, 2024) - dermatology
- Diagnostic Access Bias: Individuals in certain geographical localities have better access to medical care and, hence, may appear to have higher disease prevalence. For example, atopic dermatitis is believed to be commoner in the West – this could be due to better and earlier diagnostic facilities available than in India.19,20 17.(Chakraborty, 2024)
- Diagnostic reference test bias: These bias results when all individuals do not receive the same reference test. e.g., direct immunofluorescence studies may not be done for all patients with pemphigus vulgaris some patients

may receive only a skin biopsy-based diagnosis. It is a subtype of verification bias. Another variation of this type of bias is partial reference bias, where only some of the study participants receive the index and the reference tests.21(Chakraborty, 2024)

- Mimicry bias: When an exposure causes a disease that resembles the study disease, mimicry bias can result. For example, certain drugs are known to cause a pityriasis rosea-like reaction, which, although looks like pityriasis rosea, differs from it.28 25.(Chakraborty, 2024) - dermatology
- Unacceptable disease bias: This occurs in socially unacceptable diseases like leprosy and STDs, which result in under-reporting.30 27. (Chakraborty, 2024) - dermatology
- TODO: Other such studies were conducted in [(Fry et al., 2017)] which states that UK Biobank, a large and widely used genetic dataset, may not represent the sampling population. Researchers found evidence of a "healthy volunteer" selection bias. [150] has other examples of studies on existing biases in the data used in the medical domain. [157] also looks at machine-learning algorithms and data utilized in medical fields, and writes about how artificial intelligence in health care has not impacted all patients equally.(Mehrabi et al., 2021) –> [150] also provides an ovverview over the impact of social determinants on health, such as Economic stability, neighborhood and physical environment, education, food, community and scial context, access to healthcare and quality
- The healthy volunteer effect is a particular case: when the participants are healthier than the general population. (Delgado-Rodríguez & Llorca, 2004)

Temporal Biases

Differences in populations and their behaviour over time can lead to temporal biases (Olteanu et al., 2019). Certain studies require to track temporal data, to learn about their behaviour over time. Disease progression is also a factor measured over time (Mehrabi et al., 2021). For PASSION, temporal biases are currently irrelevant, since PASSION contains images independently of time and is not tracking the disease progression. Therefore, the listed biases in this chapter are not explained in detail, refer to the sources for further information.

Examples for temporal data biases are:

- **Longitudinal Data Fallacy** (Mehrabi et al., 2021)
- **Chronological bias** (**c9**; **c13**; Chakraborty, 2024)
- **Immortal time bias** (**c24**; **c20**; Chakraborty, 2024)

TODO: added until here

B.15.5 Algorithmic Biases

When an algorithm adds biases to unbiased input data one speaks of **Algorithmic Bias** (Baeza-Yates, 2018). This could occur due to algorithmic design choices like optimization functions, regularizations and statistically biased estimators (Danks & London, 2017).

used

- Algorithmic Bias. Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm (Baeza-Yates, 2018). The algorithmic design choices, such as use of certain optimization functions, regularizations, choices in applying regression models on the data as a whole or considering subgroups, and the general use of statistically biased estimators in algorithms (Danks & London, 2017), can all contribute to biased algorithmic decisions that can bias the outcome of the algorithms.(Mehrabi et al., 2021).

User Algorithm Interaction Biases

- **User Interaction Bias**. This biases can be triggered by the user interface or the user themselves. The user interface influences the user to behave in a certain way, which could introduce bias in the user behaviour. Users impose this (or their own) biased behavior through interaction on the algorithm (Baeza-Yates, 2018). **Presentation bias** and **Ranking bias** are further subtypes mentioned by Lerman and Hogg (2014) and Mehrabi et al. (2021).
- **Emergent Bias**. When real users interact with an algorithm, this bias arises some time after the design was completed due to changes in population. It appears more likely in user interfaces (Friedman & Nissenbaum, 1996).

Potential Biases in PASSION The user interaction biases, especially the emergent bias could potentially become an issue for PASSION, when the project starts to become publicly available teledermatology. Also, the interface design should be evaluated, so that no presentation or ranking bias gets introduced.

used

- Emergent Bias. Emergent bias occurs as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values, or societal knowledge usually some time after the completion of design (Friedman & Nissenbaum, 1996). This type of bias is more likely to be observed in user interfaces, ... This type of bias can itself be divided into more subtypes, as discussed in detail in (Friedman & Nissenbaum, 1996). (Mehrabi et al., 2021). probably less relevant at the first stage
- User Interaction Bias. User Interaction bias is a type of bias that can not only be observant on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction (Baeza-Yates, 2018). This type of bias can be influenced by other types and subtypes, such as presentation and ranking biases. (Mehrabi et al., 2021). – more relevant for later, when the application would become bigger

  - Presentation Bias. Presentation bias is a result of how information is presented (Baeza-Yates, 2018) (can only click on content they see, could be the case that user does not see all info on web) (Mehrabi et al., 2021).
  - Ranking Bias. The idea that top-ranked results are the most relevant and important will result in attraction of more clicks than others. This bias affects search engines (Baeza-Yates, 2018) and crowdsourcing applications (Lerman & Hogg, 2014).(Mehrabi et al., 2021).

External Influence Biases

Those biases can be introduced :

- **Evaluation Bias**. When inappropriate or disproprtionate benchmarks are used in model evaluation, they can introduce the benchmarks biases into the model. (Suresh & Guttag, 2021; Buolamwini & Gebru, 2018)
- **Incorporation bias**. When index tests in diagnostic accuracy studies are part of the reference tests, this results in elevated sensitivity for the index tests (**c21**; **c25**; **c26**; Chakraborty, 2024; Young et al., 2020).
- **Popularity Bias**. More popular items tend to be exposed more. Popularity metrics can be manipulated though or not reflecting good quality, this can lead to bias (Ciampaglia et al., 2018; Mehrabi et al., 2021).
- **Generalization Issues**. (**<empty citation>**) TODO: add those from young

Potential Biases in PASSION TODO: add used

- Evaluation Bias. Evaluation bias happens during model evaluation (Suresh & Guttag, 2021). This includes the use of inappropriate and disproportionate benchmarks for evaluation of applications such as Adience and IJB-A benchmarks. These benchmarks are used in the evaluation of facial recognition systems that were biased toward skin color and gender (Buolamwini & Gebru, 2018), and can serve as examples for this type of bias (Suresh & Guttag, 2021). (Mehrabi et al., 2021). – important for this thesis
- Incorporation bias: This is principally relevant for diagnostic accuracy studies when the index test forms a part of the reference test, resulting in elevated sensitivity e.g., if one wants to compare the grattage test vs. dermoscopy in psoriasis and does dermoscopy only from areas of grattage positivity, one would get a very high sensitivity for the grattage test because it was incorporated into the reference test, i.e., dermoscopy.25,26 21.(Chakraborty, 2024)
- Popularity Bias. Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation—for example, by fake reviews or social bots (Ciampaglia et al., 2018). ... this presentation

may not be a result of good quality; instead, it may be due to other biased factors. (Mehrabi et al., 2021).

B.15.6 User Biases

Cognitive Biases

Biases which are related to human perception belong to the category of cognitive biases. They are affecting how data should be presented and interpreted (Mester, 2017)

Those biases can be introduced :

- **Confirmation Bias**. When people have pre-conceptions, they will only listen to the part of presented information which reinforce those ”facts”, regardless whether the facts are true or not (Mester, 2017). In health-care, this can be observed when patients report increases in diseases due to potentially nonfactual information they found on the internet (**c15**; **c14**; Chakraborty, 2024).
- **Belief Bias**. A stronger version of the confirmation bias: Someone who is affected by this bias is so sure about their own gut feelings that they will ignore results of a data research project (Mester, 2017).
- **Previous Opinion Bias**. When performing multiple tests, the knowledge about the outcome of the previous tests probably influences the results (Chakraborty, 2024)
- **Cause-Effect Bias**. The famous senctence ”correlation does not imply causation” can be used here - when correlation between two variables is misinterpreted as a cause-effect in the wrong direction, this bias applies (Mester, 2017)
- **Historical Bias**. Preexisting biases in the world can affect the data generation process (Suresh & Guttag, 2021). Even if they reflect the current reality, it is worth to consider whether those biases should affect the algorithms in question (Mehrabi et al., 2021).
- **Content Production Bias**. User generated contents can introduce biases by systematical differences in the production process, stucture and appearance, which might stem from the users background (Olteanu et al., 2019).

Potential Biases in PASSION For PASSION, confirmation bias could lead to issues in the initial diagnosis and could therefore lead to biased data labeling. Same with the previous opinion bias. The later can be reduced when it is ensured, that the labeling experts are diagnosing the diseases independently of each other, so that they do not know the previous opinions. Cause-Effect bias is lesser an issue for PASSION, since the causes of the diseases are not analyzed. It could more be an inherit problem, that the algorithm learns wrong causes for diseases, such as appearing hair Historical bias can affect PASSIONs process in various ways. In

PASSION context, Content Production Bias could have an impact on how the images are taken.

used

-
- Cognitive bias (Mester, 2017) - statistical bias
- Previous opinion bias: In performing a second diagnostic test, if the result of a previous test is known, it is likely to influence the result. An extension of this is the Greenwald's law of lupus: the Sontheimer amendment – anything and everything that happens to a lupus erythematosus patient is correctly or incorrectly attributed to lupus.32 29. (Chakraborty, 2024) - dermatology
- Confirmation bias: This bias occurs when study participants have a preconceived notion of their disease that may not be based on facts. For example, we have observed that in North India many tinea patients report an increase in their disease due to taking meat, fish, and other so-called "hot foods". They may also present information they have collected from the internet which reinforces their beliefs.15 14.(Chakraborty, 2024) - dermatology
- Cause-effect bias (Mester, 2022; Mester, 2017) - statistical bias
- Historical Bias. Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection (Suresh & Guttag, 2021). ... search results were of course reflecting the reality, but whether or not the search algorithms should reflect this reality is an issue worth considering (Mehrabi et al., 2021) - maybe relevant
- Content Production Bias. Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contents generated by users (Olteanu et al., 2019). (Mehrabi et al., 2021) – could the quality of the pictures been related to this as well?

Behavioral Biases

Those biases can be introduced :

- **Behavioral Bias**. User behaviour can differ depending on the platforms, contexts, cultures, or datasets (Olteanu et al., 2019).
- **Self-Selection Bias**. This subtype of selection bias occurs when study participants can select themselves. Less proactive people, people with less time or interest will be excluded or underrepresented (Mester, 2022; Mehrabi et al., 2021). **Non-Responder bias** is a subtype, where part of the population is not responding e.g. to fill out a survey or post-study responses queried by postal services (Chakraborty, 2024). TODO: maybe categorize this in the data biases or the healthy volunteer bias here
- **Social Bias**. When the actions of others affect our judgment, it is called social bias. For example ratings in juries can be affected by this (Baeza-Yates, 2018).

Potential Biases in PASSION For PASSION the behavioral biases can affect who is going to the dermatologists for what reasons. Therefore, the approach to use data from different countries may be benefitial, since potentially the cultural differences could differ. Self-selection is an issue, since only those patients can be included in the database which go to the hospitals.

used

- Self-Selection Bias. Self-selection bias4 is a subtype of the selection or sampling bias in which subjects of the research select themselves. (Mehrabi et al., 2021)
- Self-selection bias - when you let the subjects of the analyses select themselves, less proactive people will be excluded TODO: could be an issue as well for PASSION, couldn't it? since the doctors probably ask the clients. One way to go is to default should be to provide access to the data. but is it ethical? (Mester, 2022; Mester, 2017)- statistical bias A variation of this is non-responder bias, where non-responders to a questionnaire differ significantly from responders.9 9. (Chakraborty, 2024) - dermatology
- Social Bias. Social bias happens when others' actions affect our judgment (Baeza-Yates, 2018). (case where we want to rate or review an item with a low score, but when influenced by other high ratings, we change our scoring thinking that perhaps we are being too harsh [(Baeza-Yates, 2018), (Wang & Wang, 2014).) (Mehrabi et al., 2021)
- Behavioral Bias. Behavioral bias arises from different user behavior across platforms, contexts, or different datasets (Olteanu et al., 2019). (Mehrabi et al., 2021) maybe, people from different countries go to the dermatologist for different diseases, based on cultural differences?

Publication Biases

Those biases can be introduced :

- **Publication Bias**
- **Hot stuff bias** is a subtype of publication bias, where Journals are less critical about trending topics, which lead to more frequent publishing of those topics. This in turn can lead to flawed meta-analyses regarding those topics (**c22**; **c23**; **c19**; Chakraborty, 2024).
- **All is Well Bias**. This bias is a different view on the hot stuff bias. Theories which align with the view of the majority are more likely to be pubhlished than an opposing view (**c7**; **c10-12**; Chakraborty, 2024).
- **Rethoric Bias**. Charismatic writing or when the press is more vocal about findings can lead to greater influence over individuals than other available facts (Chakraborty, 2024).
- **Novelty Bias**. Newer interventions appear to be better. Over time, this effect decreases (Chakraborty, 2024).

Potential Biases in PASSION These biases are relevant for all researchers. They should kept in mind when interpreting, publishing and peer-reviewing papers.

used

- Hot stuff bias: Editors of journals may be less critical about topics that are "fashionable" or currently in vogue and consequently end up publishing them more frequently, resulting in publication bias as well as hot stuff bias. It can result in flawed meta-analyses based on these studies. An example is how cutaneous manifestations of COVID-19 were published. Indian Journal of Dermatology Venereology and Leprosy stood out by choosing not to publish anything and everything related to COVID-19, thus reducing hot stuff bias.22,23 19. (Chakraborty, 2024)
- All is well bias: It is a subjective bias where theories supported by the majority tend to get more easily published than the opposing view supported by the minority. For example, ideas on the origin of endemic pemphigus supporting autoimmunity are more likely to be published than theories exploring an infectious trigger. According to some authors, this bias is very difficult to eliminate and is a variant of publication bias.10-12 7.(Chakraborty, 2024) - dermatology
- Rhetoric bias: A more charismatic piece of writing has a greater influence on the study participants than other available literature. An example is the wider use of sunscreen for polymorphous light eruption over photoprotective strategies like umbrellas, broadbrimmed hats, etc, because the lay press is more vocal about sunscreens.14 11. (Chakraborty, 2024) - dermatology
- Novelty bias: The newer an intervention, the better it appears, and with time, its efficacy seems to decrease. When ligelizumab, an IgE antagonist was first discovered, ligelizumab was believed to be better than omalizumab; however, evidence soon pointed to the contrary. 16.(Chakraborty, 2024) - dermatology

Medical Biases

Those biases can be introduced :

- **Popularity Bias**. In medicine, when more popular diseases (usually well-known or stigmatized ones) get compared with less popular diseases, clinic rates can show a distorted view. The more popular diseases appear to be over-represented over more commoner ones (**c9**; **c6**; Chakraborty, 2024).
- **Apprehension Bias**. Fear related to an upcoming procedure can lead to false evaluations, e.g. when measuring blood pressure (**c13**; Chakraborty, 2024).
- **Hawthrone bias**. Subjects might modify their behaviour when they know they are being watched. This bias can be practically utilized by introducing regular follow-ups (**c8**; Chakraborty, 2024).

- **Centripetal Bias**. Better reputations affect to which physicians or hospitals patients tend to go to. Famous specialists probably see more cases in regards of their specialty than others (**c12**; Chakraborty, 2024).

Potential Biases in PASSION PASSION must be careful in interpreting the metadata. Since the data is from hospitals, they could be biased towards more popular diseases. PASSION can potentially use Hawthrone bias to improve the work of the annotators. Centripetal bias can also be used when selecting the partners to work with.

used

- Popularity Bias: This bias arises when a particular disease is more popular (i.e. either more well-known or more stigmatised) among the participants than the disease with which it is compared. For example, if a study compares clinic attendance rates among various dermatological disorders, one would see vitiligo patients are over-represented over melasma. While melasma is commoner in the normal population, vitiligo, due to its popularity because of media publicity and other factors, tends to present earlier.9 6. (Chakraborty, 2024) - dermatology
- Apprehension bias: This results from fear and apprehensions related to an impending procedure. The classic example is the false elevation of blood pressure because the person is apprehensive of his or her blood pressure being measured.13 A variant of this is the Hawthorne bias, where subjects modify their behavior, such as regularly taking a prescribed drug or exercising, simply because they know they are being watched, but not due to any apprehensions. Hawthorne bias is practically utilised in many leprosy clinics since regular follow-up has been shown to improve adherence to therapy based on Hawthorne bias. 8. (Chakraborty, 2024) - dermatology
- Centripetal bias: Patients tend to go to more reputed physicians and hospitals than others. For example, a famous or better-known cosmetologist with a good reputation tends to see more cases than other cosmetologists. 12.(Chakraborty, 2024) - dermatology

Appendix Bibliography

Baeza-Yates, R. (2018). Bias on the web. *Commun. ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581
Mehrabi 9.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification [ISSN: 2640-3498]. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. Retrieved March 16, 2025, from https://proceedings.mlr.press/v81/buolamwini18a.html
Mehrabi 24, demographic (skin type and gender).

Chakraborty, A. (2024). Biases in dermatology: A primer [Publisher: Scientific Scholar]. *Indian J Dermatol Venereol Leprol*, *90*(2), 250–254. https://doi.org/10.25259/IJDVL_126_2023
0 citations (but from 2024), list of lots of biases.

Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality [Publisher: Nature Publishing Group]. *Sci Rep*, *8*(1), 15951. https://doi.org/10.1038/s41598-018-34203-2
Mehrabi 117.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research [Publisher: SAGE Publications Ltd]. *Conflict Management and Peace Science*, *22*(4), 341–352. https://doi.org/10.1080/07388940500339183
Mehrabi 38, difficultis regarding ommitted variable and overcoming methods.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654
Mehrabi 44.

Delgado-Rodríguez, M., & Llorca, J. (2004). Bias [Publisher: BMJ Publishing Group Ltd Section: Continuing professional education]. *Journal of Epidemiology & Community Health*, *58*(8), 635–641. https://doi.org/10.1136/jech.2003.008466

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Trans. Inf. Syst.*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561
Mehrabi 53.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., & Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the

general population. *American Journal of Epidemiology*, *186*(9), 1026–1034. https://doi.org/10.1093/aje/kwx246
Mehrabi 54.

Hargittai, E. (2007). Whose space? differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication*, *13*(1), 276–297. https://doi.org/10.1111/j.1083-6101.2007.00396.x
Mehrabi 64.

HP, S. (2022, November 1). *Sampling — statistical approach in machine learning* [Analytics vidhya]. Retrieved March 28, 2025, from https://medium.com/analytics-vidhya/sampling-statistical-approach-in-machine-learning-4903c40ebf86

Lerman, K., & Hogg, T. (2014). Leveraging position bias to improve peer recommendation [Publisher: Public Library of Science]. *PLOS ONE*, *9*(6), e98914. https://doi.org/10.1371/journal.pone.0098914
Mehrabi 93.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning [Publisher: ACMPUB27New York, NY, USA]. *ACM Computing Surveys (CSUR)*. https://doi.org/10.1145/3457607

Mester, T. (2017, August 28). *Statistical bias types explained - part2 (with examples)* [Data36]. Retrieved March 22, 2025, from https://data36.com/statistical-bias-types-examples-part2/

Mester, T. (2022, May 16). *Statistical bias types explained (with examples) - part1* [Data36]. Retrieved March 8, 2025, from https://data36.com/statistical-bias-types-explained/

Montoya, L. N., Roberts, J. S., & Hidalgo, B. S. (2025). Towards fairness in AI for melanoma detection: Systemic review and recommendations. In K. Arai (Ed.), *Advances in information and communication* (pp. 320–341). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-84460-7_21
2025.

Mustard, D. B. (2003). Reexamining criminal behavior: The importance of omitted variable bias. *The Review of Economics and Statistics*, *85*(1), 205–211. https://doi.org/10.1162/rest.2003.85.1.205
Mehrabi 114.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries [Publisher: Frontiers]. *Front. Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00013
Mehrabi 120.

Riegg, S. K. (2008). Causal inference and omitted variable bias in financial aid research: Assessing solutions [Publisher: Johns Hopkins University Press]. *The Review of Higher Education*, *31*(3), 329–354. Retrieved March 16, 2025, from https://muse.jhu.edu/pub/1/article/232773
Mehrabi 131.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017, November 22). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. https://doi.org/

10.48550/arXiv.1711.08536
Mehrabi 142Comment: Presented at NIPS 2017 Workshop on Machine Learning for the Developing World.

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305
Mehrabi 144.

Wang, T., & Wang, D. (2014). Why amazon's ratings might mislead you: The story of herding effects [Publisher: Mary Ann Liebert, Inc., publishers]. *Big Data*, *2*(4), 196–204. https://doi.org/10.1089/big.2014.0063
Mehrabi 151.

Young, A. T., Xiong, M., Pfau, J., Keiser, M. J., & Wei, M. L. (2020). Artificial intelligence in dermatology: A primer. *Journal of Investigative Dermatology*, *140*(8), 1504–1512. https://doi.org/10.1016/j.jid.2020.02.026
209 citations.

C Fairness Metrics

According to Mehrabi et al. (2021), fairness can be achieved on a group level, subgroup level or even for an individual. Group fairness is about treating different groups as equal. Individual fairness tries to achieve similar predictions for similar individuals. Subgroup fairness tries to incorporate the best properties of the other two levels to improve the outcome in larger collections of subgroups (Mehrabi et al., 2021).

Table C.1 shows the list of fairness definitions, structured in those categories.

Fairness Definitions	**Mentioned in Context of**	
	ML	**Dermatology**
Group Fairness		
Conditional Statistical Parity	X	
Demographic/Statistical Parity	X	
Equal Opportunity	X	
Treatment Equality	X	
Test Fairness	X	
Equalized Odds	X	
Subgroup Fairness		
Subgroup Fairness	X	
Individual Fairness		
Counterfactual Fairness	X	
Fairness Through Awareness	X	
Fairness Through Unawareness	X	
Not Categorized		
Fairness in Relational Domains	X	

Table C.1: Fairness definitions based on Mehrabi et al. (2021)

The specific fairness definitions can be found in Mehrabi et al. (2021). In general, they try to get similar probability outcomes for 'unprotected' or 'protected' groups. This list summarizes how they work:

- *Demographic/Statistical Parity* and *Conditional Statistical Parity*: The parity checks that the likelihood of a positive outcome is the same for both protected groups (Dwork et al., 2012; Mehrabi et al., 2021). The conditional version adds legitimate factors before calculating the statistical parity (Corbett-Davies et al., 2017).

- *Equalized Odds*, *Test Fairness*, and *Equal Opportunity*: In all these methods, protected and unprotected groups should have equal rates of positive outcomes when belonging to the positive class. These methods essentially compare the groups' TPRs. *Equalized Odds* is a more restrictive since it also checks for similar false positive rates (Mehrabi et al., 2021; Verma & Rubin, 2018).
- *Treatment Equality*: It compares the false negative and false positive rates (Wang & Wang, 2014)
- *Counterfactual Fairness*: This approach is different from the others as it is testing the same individual in both different demographic groups with the intention that the outcome is the same (Kusner et al., 2017; Mehrabi et al., 2021). It differs from the first group of fairness metrics since it does not compare the likelihoods of the outcomes for any person in a group, but checks how the exact same individual would be treated if it was in the other group.
- *Fairness Through Awareness*: This method compares similar individuals based on similarity metrics to get a similar outcome (Dwork et al., 2012; Mehrabi et al., 2021)
- *Fairness Through Unawareness*: This measure is ensuring that protected attributes are not explicitly used in decision-making (Grgic-Hlača et al., 2016; Kusner et al., 2017).
- *Fairness in Relational Domains*: This notion also takes into consideration relational structures between individuals (Farnadi et al., 2018).

D PASSION Dataset Distribution Analysis

The data in Table D.1 shows the distribution of the values of the individual metadata attributes in the PASSION dataset. The data has been generated with a python script TODO: add/refer to python script. In Figure D.1, the data is visualized.

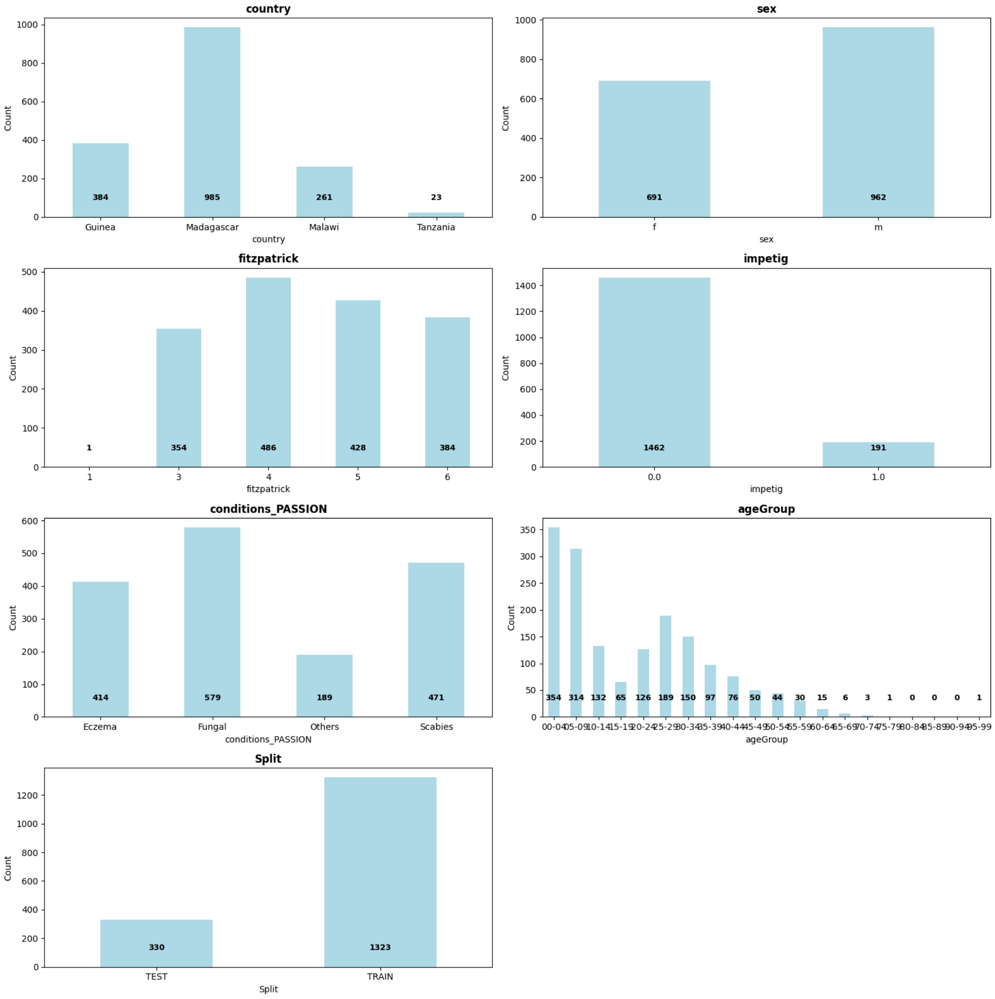


Figure D.1: PASSION dataset distribution analysis on group level

Split	Column	Value	Count	Percent
PASSION_split	country	Guinea	384	23.23
PASSION_split	country	Madagascar	985	59.59
PASSION_split	country	Malawi	261	15.79
PASSION_split	country	Tanzania	23	1.39
PASSION_split	sex	f	691	41.8
PASSION_split	sex	m	962	58.2
PASSION_split	fitzpatrick	1	1	0.06
PASSION_split	fitzpatrick	3	354	21.42
PASSION_split	fitzpatrick	4	486	29.4
PASSION_split	fitzpatrick	5	428	25.89
PASSION_split	fitzpatrick	6	384	23.23
PASSION_split	impetig	0.0	1462	88.45
PASSION_split	impetig	1.0	191	11.55
PASSION_split	conditions_PASSION	Eczema	414	25.05
PASSION_split	conditions_PASSION	Fungal	579	35.03
PASSION_split	conditions_PASSION	Others	189	11.43
PASSION_split	conditions_PASSION	Scabies	471	28.49
PASSION_split	ageGroup	00-04	354	21.42
PASSION_split	ageGroup	05-09	314	19.0
PASSION_split	ageGroup	10-14	132	7.99
PASSION_split	ageGroup	15-19	65	3.93
PASSION_split	ageGroup	20-24	126	7.62
PASSION_split	ageGroup	25-29	189	11.43
PASSION_split	ageGroup	30-34	150	9.07
PASSION_split	ageGroup	35-39	97	5.87
PASSION_split	ageGroup	40-44	76	4.6
PASSION_split	ageGroup	45-49	50	3.02
PASSION_split	ageGroup	50-54	44	2.66
PASSION_split	ageGroup	55-59	30	1.81
PASSION_split	ageGroup	60-64	15	0.91
PASSION_split	ageGroup	65-69	6	0.36
PASSION_split	ageGroup	70-74	3	0.18
PASSION_split	ageGroup	75-79	1	0.06
PASSION_split	ageGroup	80-84	0	0.0
PASSION_split	ageGroup	85-89	0	0.0
PASSION_split	ageGroup	90-94	0	0.0
PASSION_split	ageGroup	95-99	1	0.06
PASSION_split	Split	TEST	330	19.96
PASSION_split	Split	TRAIN	1323	80.04

Table D.1: Distribution of metadata attributes in the PASSION dataset

TODO: check the gls all unused.