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 in Dermatology 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

First and foremost, I would like to express my gratitude to Ludovic Amruthalingam for giving me the opportunity to work on this thesis and for his dedicated supervision throughout. His support, feedback, and guidance were invaluable to the successful completion of this work.

Many thanks go to Philippe Gottfrois, the main author of the PASSION project, for sharing valuable information and dermatology-specific insights, which was essential for my work.

I also thank Simone Lionetti for being ready to step in as deputy supervisor, helping with setting up the initial assignment and providing feedback.

As a LaTeX beginner, I am grateful to Pascal Baumann for his assistance, which made the learning process significantly smoother.

I would also like to acknowledge the broader research community working towards fairer and more accessible AI. Their efforts not only laid the groundwork for the concepts and metrics explored in this thesis, but contribute to address disparities in the current world. They inspire others for responsible innovation in machine learning.

Last but not least, I would like to thank my boyfriend, friends, family, and coworkers for their encouragement, patience, and emotional support. Special thanks to my mum and sister for their careful proofreading and helpful suggestions.

Your support made this thesis possible.

Nadja Stadelmann, 2025

TODO: reenable in the end

AI Usage Declaration

To write this bachelor thesis, OpenAI's language model ChatGPT (free version, GPT-3.5) to support specific tasks throughout the research and writing process. The assistance provided mainly included:

- Suggesting improvements for phrasing technical descriptions and argumentation clearly while preserving original content.
- Suggesting structures in the report.
- Summarizing text passages.
- Refining LaTeX formatting and ensuring structural consistency.
- Create initial code samples to remove boilerplate work during development.

The core research contributions and evaluations were carried out independently.

Abstract

Skin diseases affect up to 80% of children and adolescents in Sub-Saharan Africa, yet dermatology treatment remains often inaccessible, with fewer than one dermatologist per million inhabitants. AI-supported teledermatology can offer a potential solution, for example by enabling early triage of skin conditions. However, current AI models often perform poorly on highly pigmented skin due to demographic biases in current dermatology datasets, which predominantly contain images of low pigmented skin. The PASSION project addresses this gap by building a dataset focused on patients with highly pigmented skin in Sub-Saharan Africa to reduce demographic bias in dermatology AI models.

This thesis evaluates the effectiveness of bias mitigation strategies in the PASSION context. A structured methodology was followed, including a contextualized literature review on biases, fairness and mitigation methods in medical AI, analysis of metadata completeness and representation of demographic groups in the PASSION dataset, and development of a reproducible fairness assessment pipeline. Subgroup fairness was analyzed using equalized odds, a metric that accounts for both true and false positives, which is crucial to detect systematic diagnosis errors. Stratified splitting was applied as a mitigation method to validate the newly established fairness assessment pipeline.

Reproducing PASSION's results involved resolving metadata linkage issues, including outdated file references that required manual correction. Due to time and resource constraints, the focus shifted from testing multiple mitigation methods to establishing a robust fairness assessment pipeline and performing initial subgroup fairness evaluations. Custom implementations were required to address limitations in Fairlearn's multiclass support. While the evaluation's statistical robustness was limited by single-run experiments, the results offer valuable first insights. Furthermore, several limitations in the available metadata were identified, including the absence of socioeconomic status, clinic details, and image quality information. Those dataset limitations, along with the limited representation of several subgroups, reduced the scope of the bias evaluation.

The thesis provides a categorized overview of bias types relevant to PASSION, identifies suitable mitigation strategies, and delivers a working fairness assessment pipeline. This pipeline was used to evaluate the ResNet50 and ResNet18 models, with initial fairness and performance results obtained. Evaluating stratified splitting demonstrated insights into its impact on fairness, while also testing the assessment pipeline. Also, concrete recommendations to improve the fairness in PASSION were provided. The developed code will be integrated into the project's GitHub repository.

The work serves as a foundation for future bias detection and mitigation in PASSION. Future work is essential to build upon it. The efforts should improve metadata, adding more diverse data, potentially combining the dataset with other dermatology datasets, repeat experiments for statistical robustness, and extend the pipeline into a complete bias assessment tool for dermatology AI.

Contents

Todo list

TODO: also solve todos in the code ;)

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

TODO: integrate missing citations

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 (Fairlearn contributors, n.d.). xii, 25,

equalized odds ratio The ratio of true positive and false positive rates between subgroups, used as a group fairness metric (Fairlearn contributors, n.d.). xii, 25,

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 (Fairlearn contributors, n.d.). 12, 19, 25, 26, 27, 30, 31, 39, 40

Fitzpatrick skin type A skin classifier based on the skins' reaction to ultraviolet light, developed by dermatologist Dr. Thomas Fitzpatrick (Gottfrois et al., 2024). xii, 3,

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. 17, 19

Jupyter Notebook Executable files, often used in ML to write Python code and add explanations in text form. 5, 28, 54

pediatric A medical term for infants, children and adolescents (Farlex, n.d.). 1, 3, 35

proxy variable "one or more variables that encode the protected attribute with a substantial degree of accuracy" (S. Wang et al., 2021). 15, 18, 28, 33, 34, 61

teledermatology dermatological care from a distance, supported by modern technology (Pala et al., 2020). 1, 4, 14, 17, 23, 55, 65, 66, 69, 70, 73

Acronyms

AI Artificial Intelligence. 1, 3, 7, 9, 10, 14, 17, 18, 22, 23, 55, 56, 58, 59, 68

EOD equalized odds difference. *Glossary:* equalized odds difference, 25, 26, 30, 32, 41

EOR equalized odds ratio. *Glossary:* equalized odds ratio, 25, 26, 30, 32, 47

FPR false positive rate. 11, 31, 32, 38

FST Fitzpatrick skin type. *Glossary:* Fitzpatrick skin type, x, 3, 4, 5, 18, 26, 30, 33, 34, 35, 38, 39, 41, 46, 47, 54, 57, 58, 59, 61, 64, 73

HSLU Lucerne University of Applied Sciences and Arts. xi, 17

ML Machine Learning. ix, x, xi, 3, 4, 7, 8, 9, 10, 11, 13, 23, 27, 46, 47, 56, 59, 65, 67, 80, 82, 83, 84

TPR true positive rate. 11, 31, 32, 38, 80

1 Problem Statement

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 images 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 has shown, that patients with a lower socioeconomic status have more advanced disease progression at the time of diagnosis, which can lead to different appearances of the same disease (British Association of Dermatologists (BAD), 2021). Since the AI models use images as input and can only learn to diagnose diseases based on their appearances in the images, factors affecting 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.

2 State of Research

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, covering its medical scope and technical components.

While the overall objective is to enhance the accessibility of dermatological care by developing fair and inclusive AI systems, PASSION focuses on prevalent 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 influence the feasibility of the mitigation methods that can 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 later on. 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.

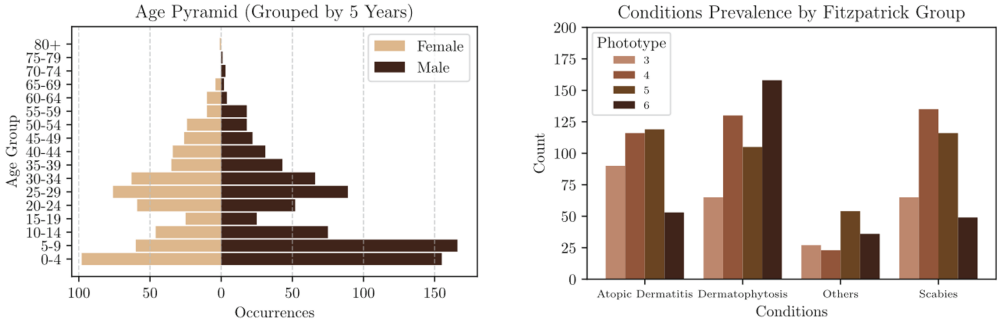

Fig. 1. Age distribution per gender.

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 images, 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. 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 (Mehrabi et al., 2021). 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 Other Types of Biases

The literature describes numerous types of bias. Over 50 were identified during this research. These biases were grouped into categories to provide an overview.

Table 2.2 provides an overview of the types of biases identified in the literature. The base categorization follows the main stages of the ML lifecycle where the observed bias in the model originated (data collection, algorithm design, and user interaction) as proposed by Mehrabi et al. (2021).

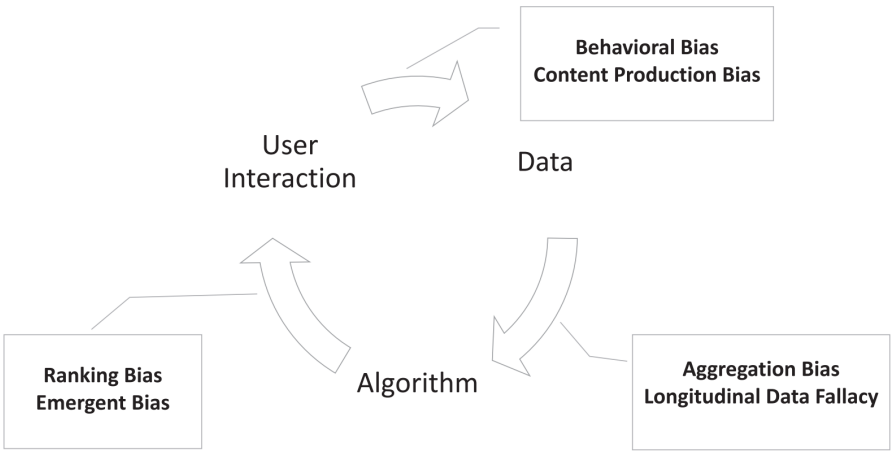

Figure 2.2: ML lifecycle with fitting biases (Mehrabi et al., 2021).

Each specific category groups together similar kinds of biases. Notably, some biases may reasonably fit into multiple categories. Definitions of the categories and the specific biases, can be found in appendix B List of Biases.

Bias	Mentioned in Context of	
	ML	**Dermatology**
Data Collection		
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 Biases	X[1]	X[4]
Algorithmic Design		
Algorithmic 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 2.2: Bias categories - grouped according the ML lifecycle of Mehrabi et al. (2021)

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 2.3 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	Mentioned in Context of Dermatology
Related to Disease Presentation		
Skin Type	X[1,2,7]	X[12,13]
Skin Undertones		X[13]
Socioeconomic 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]	
Socioeconomic 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 2.3: 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 2.4.

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 2.4: 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 2.3, the model satisfies equalized odds, and fairness is achieved. Since equalized odds compares conditional probability distributions across groups, it is a group fairness metrics.

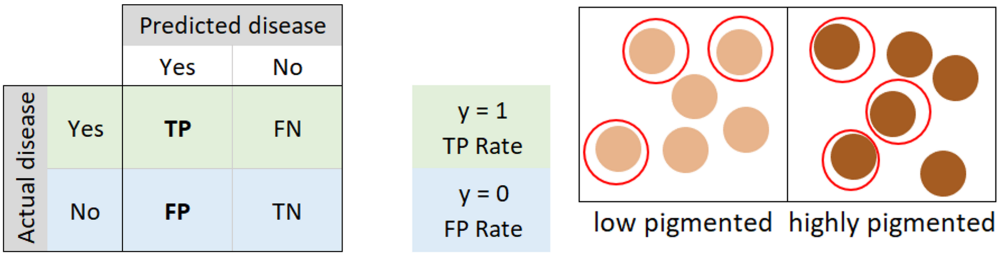

Figure 2.3: Equalized odds mechanics, inspired by Kearns et al. (2019).

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 (Fairlearn 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 2.3) but still exhibit substantial disparities within subgroups, such as older individuals with darker skin tones (Figure 2.4) (Kearns et al., 2018, 2019).

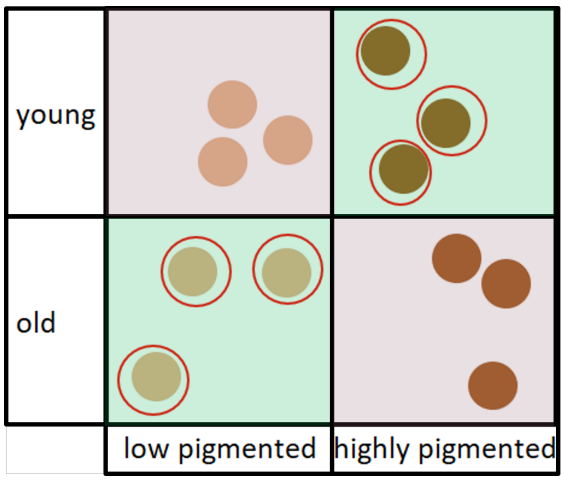

Figure 2.4: Equalized odds violations on subgroups, inspired by Kearns et al. (2019).

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

When biases are identified, they can be reduced through various mitigation methods. Since biases may arise at different stages of the ML life cycle, corresponding mitigation strategies exist for each stage. These include e.g., methods for reducing representation disparities in datasets (pre-processing), approaches targeting model architecture and algorithm design (in-processing), and techniques aimed at interpreting model outputs (post-processing) (Mehrabi et al., 2021).

This thesis provides an overview of the mitigation methods currently recognized in the field of ML, as AI engineers must understand the available methods for reducing bias (Mehrabi et al., 2021). Table 2.5 presents a broad categorization of these methods. The focus lies on approaches related to fair data collection and design, as well as fair classification, given the nature of the ML task at hand. Specific examples for each category are listed in Appendix D; for further details, the original sources should be consulted.

Mitigation Method Categories	**Mentioned in Context of**	
	ML	**Dermatology**
Fair Data Collection and Design		
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]
Removing Sensitive Attributes	X[1]	
Fair Classification		
Satisfy Fairness Definitions	X[1]	
Algorithmic Adaptions for Fairness	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 Tasks		
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 2.5: Mitigation method categories

3 Ideas and Concepts

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 consist 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 be 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 be assessed relative to the baseline.

3.2 PASSION Dataset Assessment

In order to determine 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 assessment of representativeness can be postponed until the model output indeed proofs to be biased.

Furthermore, the available metadata determines which biases can be 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 regarding:

- 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

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 process model used, how the progress and risk are managed, and the technical constraints available. This gives an overview of the constraints and the general plan of this thesis.

4.1.1 Process Model

The project follows the waterfall model. This means the tasks are completed 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 while keeping the project management overhead small. This project is separated into 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 based on 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 of existing research. Then, more detailed papers on dermatology AI are conducted to gain further insight into 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 within the context of the PASSION project. This involves analyzing the findings from the literature review in terms of their relevance to teledermatology and similar healthcare applications. This analysis considers 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. Also, crucial demographic biases are included as high, since PASSION's aim is to create more accessible AI models.
- **Medium.** Generally relevant to diagnostic AI and PASSION, but do not seem to have the biggest impact towards PASSION's primary goals.
- **Low.** Related to PASSION but only limited. E.g., in the far future it could potentially impact PASSION.

- **Not Applicable.** Not relevant for PASSION due to fundamental differences in domain, type of data, or type of model. For this categorization, no mitigation methods are listed.

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.

To test the fairness assessment which will be implemented, a mitigation method is selected that could feasibly be implemented within the given timeframe. The selection excluded any approaches requiring new data collection and focused on techniques applicable to the existing dataset and resources due to the time limit of this thesis.

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 is conducted using the original project setup. The fairness metric selected in subsection 5.1.3 is implemented in a fairness assessment pipeline 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. The Fairlearn library is used whenever possible to rely on standard implementations. 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 attributes 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

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

The categorization of biases in Appendix B has been enhanced by indicating how relevant they are to the PASSION context. Also, mitigation strategies mentioned in the research and based on the contextualization have been added.

Among the categories, *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.

Here, the most interesting biases in regards of PASSION are highlighted. An extensive list of the identified types of bias is provided in appendix B List of Biases and will be shared with the PASSION team for further evaluation. Appendix B also provides some further insight into potentially mitigating the following biases.

For completeness, it is also notable that certain biases can be used to improve AI models or the surrounding research. E.g., the *Hawthrone bias* as explained in subsection B.14.3 could improve a projects quality by introducing monitoring via 4-eyes reviews of code or the data annotating process.

Annotator Bias

This 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). In PASSION, annotator bias could particularly affect the labeling of skin tones, which are somewhat dependent on individual perception. This bias could further lead to inconsistent classifications of skin conditions across different demographic groups.

Aggregation Bias

This 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 sex or ethnicity) are not properly accounted for (Mehrabi et al., 2021; Suresh & Guttag, 2021). Aggregation bias is a significant concern in PASSION, since it involves multiple sensitive demographic factors which impact skin disease prevalence and appearance. The model needs to account for these factors to avoid generalized conclusions that might harm certain subgroups.

Image Quality Bias

This bias occurs when the quality of an image, such as the zoom level, focus, lighting, or even different hardware affects how a ML model classifies or diagnoses the image. Poor image quality can lead to misclassification or lower prediction accuracy (Young et al., 2020). Since PASSION will be used in a teledermatology context, it will not be feasible to fully standardize image acquisition which was also proposed by Young et al. (2020). The image quality assessment proposed above is probably the best method going forward. Also, the biased outcome regarding the countries could be an indicator, that this bias indeed exists in PASSION. *Visual artifacts bias* is a related bias which occures based on e.g., the presence of hair in the data and also could be influenced by the data gathering process.

Previous Opinion Bias

When the knowledge of prior results or diagnoses influences the interpretation of new data, leading to biased conclusions (Chakraborty, 2024). This is not only relevant to PASSION's labeling process but even more importantly, it also affect real-world diagnoses once PASSION is deployed. For example, if both the patient and the dermatologist are aware of the model's prediction before the dermatologist evaluates the case, the model's output could influence the final diagnosis. Therefore, it is crucial that the model's prediction is not shown to users - at least not during triage and prior to the clinical assessment - unless it concerns a condition that users can safely treat themselves.

Diagnostic Access Bias

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 (Chakraborty, 2024). PASSION addresses diagnostic access bias in dermatology AIs regarding Sub-Saharan Africa. However, the bias could still be relevant in the dataset, depending on which clinics were chosen for data selection.

5.1.2 Sensitive Features

Some of the listed features in Table 2.3 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 features do not seem to be further relevant for PASSION.

5.1.3 Fairness Metrics

This chapter focuses on those fairness metrics which can 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 2.4, 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 image. 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.

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 (Fairlearn 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). A similar approach is proposed 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).

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 (Fairlearn 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 (Fairlearn contributors, n.d.).

5.1.4 Mitigation Methods

During the research phase, mitigation methods were broadly categorized as outlined in section 2.4. Due to time constraints, a detailed relevance assessment of these methods for PASSION was not carried out. Nonetheless, Appendix B offers an initial interpretation of potentially relevant mitigation strategies, based on the contextualization of identified biases. These interpretations serve as a starting point for selecting appropriate techniques from the established summary and guide the prioritization of next steps. Furthermore, this chapter provides additional insights into mitigation methods that seem immediately relevant.

Relevant Mitigation Methods for PASSION

Based on prior findings, several mitigation methods already included in PASSION were identified by reviewing the source code and publication of Gottfrois et al. (2024):

- **Metadata exclusion during training**, corresponding to *blinding* as discussed by Chakraborty (2024).
- **Stratified splitting** of the dataset, referenced in F. Chen et al. (2024). However, it is unclear whether sensitive attributes where used in the splitting.
- `StratifiedKFold` is used to maintain stratification during cross-validation. Notably, this method stratifies only on the target labels (scikit-learn developers, n.d.).
- Performance metrics include **balanced accuracy and macro F1**. These metrics are particularly relevant for fairness-aware ML, as they assign equal weight to each class which mitigates the effect of class imbalance in the performance report. TODO: Add citation from ML course
- **Image augmentation using color jitter** was tested but later discarded due to performance degradation (Gottfrois et al., 2024).

To highlight further mitigation strategies that may be relevant for PASSION, the following list outlines potential next steps:

- Reevaluate the existing PASSION split and explore the inclusion of additional metadata in the stratification.
- Extend the current K-fold cross-validation to also incorporate stratification using metadata attributes.
- Revisit image pre-processing. For example, Hameed et al. (2020) suggest converting images to grayscale as a pre-processing step for hair removal.
- Apply a *price of fairness* to integrate fairness constraints directly into model training. This concept has been introduced in fair regression (**M14_**) and in fair classification tasks by **M12_<empty citation>**.
- Use fairness-aware model frameworks such as Fairlearn to guide algorithmic decisions (**M155_**).
- Explore preferential sampling to enhance the representation of underrepresented groups (**M75_**). However, since this approach can resemble oversampling, which was cautioned against in expert discussions TODO: cite midterm-presentation, a more sustainable approach might be extending the dataset with samples from more diverse populations, as described in the *Data Gathering* section of Table D.1.

Selected Method for This Thesis

Given the outlined constraints, this thesis focuses on enhancing the existing stratified splitting method as a concrete mitigation approach. This method can be applied to the PASSION dataset and compared to the current baseline to test the effectiveness of the fairness assessment framework introduced in section 4.6.

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 attributes are not mentioned in the paper, the analysis for proxy variables was more difficult than expected. The result is based on the sensitive features and biases mentioned in the literature.
  Also, what the country attribute represents in PASSION is not entirely clear based on the documentation. To clarify its meaning, the main author of PASSION was contacted. Recommendations for more precise alternatives were provided to be checked by the PASSION team for all attributes which appear to potentially be used as a proxy variable.
- **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. This increased the time effort for the detailed analysis. The script is part of the newly created `evaluator` class and is designed to run independently. 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 was done manually for now, since there were not too many values. Note that this script currently provides the distribution of available cases. For an even more in-depth view, the script should be enhanced to also report the amount of images.
  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 did not seem to 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 the model to be reloaded, 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 incorrect 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 *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, 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 were combined with custom implementation to compute 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 (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 were 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.

To compare models, the aggregated values must 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 subset.

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 hinders stratification since at least two samples per subgroup are required for an 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 a total of 12 models. The seeds were fixed to ensure compatibility. Unfortunately, the seed was mistakenly changed between generating the different strategies. Therefore, the models were evaluated separately 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 conducted entirely based on EOD, since EOR mostly reported values close to zero for 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 issue by collecting more subgroup-specific data.

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

6 Evaluation and Validation

The initial objective of this thesis was to evaluate the effectiveness of mitigation methods for reducing demographic biases in the PASSION dataset, to support more accessible dermatological care and ultimately improving patients' lives.

As the thesis progressed, the focus shifted from the application and comparison of multiple mitigation strategies to establishing a structured overview of known biases. The thesis laid the foundation for future mitigation planning tailored to PASSION, thus provide a crucial step towards the inital goal. The scope further evolved into designing a fairness assessment pipeline based on current research practices and supported by available Python libraries. The use of stratified splitting as a mitigation strategy served as a case study to test and validate this pipeline.

Overall, the fairness assessment presented in this thesis provides a valuable starting point and a practical blueprint for identifying and addressing biases in PASSION. However, to draw statistically significant and robust conclusions, the pipeline must be further enhanced, as discussed in this chapter.

6.1 PASSION Dataset Assessment

This section describes the PASSION dataset assessment results. 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 5-year intervals, 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 dermatological conditions)

The attribute *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 patient's *geographic location* or residence. 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, to mitigate medical biases and ascertainment bias, more clinics should be included into the data collection process. Then, the clinic and more detailed data about it should be added to the dataset. As for the image quality, it would further improve the dataset by tackling image biases if the information would be quantified through e.g., the used phone and camera settings. Currently, the only indication of different image gathering process is hidden in the country attribute. 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 the sources of related biases, the suggested changes to the metadata would need to be introduced. 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 E 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, these 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 more closely, to see if their representation in the dataset must be

addressed further. TODO: cite https://academic.oup.com/bjd/article-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%. No 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 these imbalances significantly affect model fairness, the dataset composition may need to be revised.

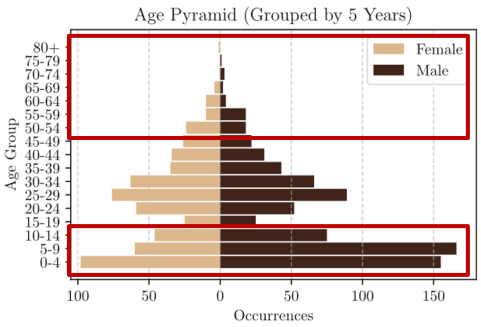

Fig. 1. Age distribution per gender.

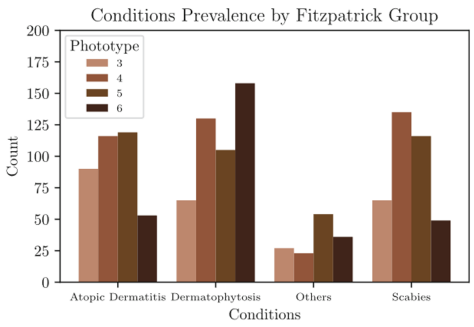

Fig. 2. Prevalence per FST.

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

Even though this analysis does not take into consideration multiple images per case, these findings highlight representation disparities across several demographic and clinical factors. Such disparities should be considered during training and when evaluating fairness, especially when assessing subgroup-specific performance.

It is important to note that the analysis provided only is a high-level overview at the group level. Detailed subgroup representation has not yet been assessed in detail. 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. TODO: consider instead: The baseline fairness performance of `ResNet50` and `ResNet18` was evaluated to establish a reproducible reference point for assessing the effectiveness of subsequent fairness mitigation strategies.

Only minor performance differences were observed between the two model variants (Table 6.1), and the subgroup-level fairness patterns remained largely consistent (Table 6.2). As a result, ResNet18 was considered a suitable substitution model for the experiments conducted in this thesis.

Detailed findings are presented in the following subsections.

Metric	ResNet18	ResNet50
Overall		
Accuracy	0.69	0.71
Macro F1	0.69	0.71
Weighted F1	0.69	0.71
Balanced Accuracy	0.69	0.71
Class-Level F1 Scores		
Eczema	0.60	0.65
Fungal	0.69	0.72
Others	0.73	0.72
Scabies	0.73	0.74

Table 6.1: Baseline model performance comparison: ResNet18 vs. ResNet50.

Metric - Subgroup Mean Compared To Overall	ResNet18	ResNet50		
Overall				
avg	0.49	0.49		
best	0.03	0.02		
worst	0.73	0.74		
median	0.54	0.57		
std. dev. sub pop.	0.24	0.23		
std. dev. whole pop.	0.23	0.22		
Avg. Per Subgroup			**# Sub-groups**	**Min Support**
fitzpatrick	0.10	0.12	4	87
sex	0.03	0.02	2	425
ageGroup	0.46	0.48	14	2
country	0.34	0.37	4	19
fitzpatrick, sex	0.18	0.17	8	32
fitzpatrick, ageGroup	0.67	0.63	47	1
fitzpatrick, country	0.51	0.55	11	2
sex, ageGroup	0.56	0.60	26	2
sex, country	0.38	0.42	8	4
ageGroup, country	0.73	0.60	30	1
fitzpatrick, sex, ageGroup	0.70	0.74	81	1
fitzpatrick, sex, country	0.54	0.57	19	1
fitzpatrick, ageGroup, country	0.73	0.67	62	1
sex, ageGroup, country	0.73	0.74	55	1
fitzpatrick, sex, ageGroup, country	0.73	0.74	102	1

Table 6.2: Baseline fairness assessment comparison: ResNet18 vs. ResNet50.

6.3.1 Bias in the Baseline

TODO: link output Subgroup fairness results revealed slight disparities but the trends are consistent. The summary of Table 6.2 highlights the following insights into existing biases:

- **Sex:** Only slight disparities indicated.
- **FST:** Indications of existing biases, accompanied by an uneven support distribution.
- **Age Group:** This attribute holds the largest disparities among the main categories. Some age groups had very low support (down to 0, the lowest existing group had 2).
- **Country:** Shows fairness disparities and skewed representation.
- **Intersectional Analysis:** Bias increase across more granular subgroup combinations. Even for attributes with lower individual disparities (e.g., FST, sex), their combinations uncovered additional bias patterns. With more intersected dimensions, the disparities increased significantly.

To gain deeper insights, the data was evaluated on subgroup level in an exploratory analysis:

- **Sex:** Both models showed a slight bias toward women, with higher TPR for female patients.
- **FST:** Both models consistently privileged FST V and underprivileged FST VI. Notably, FST III and VI showed different behavior across models, being more privileged in the big model.
- **Age:** The impact of age was relatively small on the FPR but showed large disparities in the TPR in both models. The age groups 0–14 and 25–29 were generally better off, whereas groups 20–24 and 30–69 were more often underprivileged. No samples for the 70+ group were available in the test data.
- **Intersectional Analysis:** 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:** In both models, 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 scripts provided enable further detailed analysis and subgroup comparisons.

These findings reinforce the importance of evaluating fairness not only at group level but also across subgroups, as emphasized by Kearns et al. (2018, 2019). As discussed in subsection 5.5.1, fairness metric volatility tends to increase when subgroup sample sizes are low. Collecting more data is thus essential to ensure robust assessments.

In the mean time, one potential mitigation strategy involves consolidating age groups, e.g., using an aggregation often used in dermatology. As neighboring age ranges often show similar behavior, merging them may help improve subgroup support without significantly hiding underlying fairness trends. This consolidation would also facilitate clearer fairness evaluations by reducing metric noise from very small subgroups.

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 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 identified disparities among subgroup, supporting the idea that fairness analysis among subgroups is important for reducing biases in dermatology models, including PASSION.

6.3.4 Aggregation Trade-offs

Although aggregating fairness metrics at subgroup and model levels provides helpful summaries, it 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 negatively impact the overall model performance.

These results should be interpreted with caution, given that the experiment faced several limitations. To achieve more statistically robust findings, additional data is required. 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 could inform future data collection efforts to achieve fairer outcomes.

6.4.1 Demographic Representation Across 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 that 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 (see Table 6.3). Similarly, FST IV and V are overrepresented in training data (Table 6.4), which may contribute to the observed bias. However, FST VI is evenly distributed across the subsets, yet model performance remains poor. This suggests that data imbalance is not necessarily the sole cause of observed biases. Nevertheless, 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 6.3: PASSION Dataset: Sex distribution (train, test, overall).

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

6.4.2 Initial Training

As shown in Table 6.5, the fairness assessment of the initial model training indicated that, for strategy A (placing single records in training data), 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 6.6), where singletons were put in the validation data, the fairest model was achieved by stratifying only on the target labels, due to the following:

- 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 6.5: 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 6.6: Stratified Split: Fairness summary (seed 32, single-record validation stratification).

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

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

6.4.3 Cross-Validation

The results of the 5-fold cross-validation step confirmed that the fairest models for both single-record handling strategies were consistently produced by including *country* and *fitzpatrick* in the stratification (see Table 6.7, Table 6.8).

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 6.7: 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 6.8: Stratified Split: Fairness summary (5-fold CV, seed 42, training stratification).

6.4.4 Baseline Comparison

The final evaluation on the original test set (Table 6.9) confirms that applying stratified splitting impacts model fairness, especially at subgroup levels. Stratifying on *country* and *fitzpatrick* improved fairness, especially when single records were put in the training set. For the other strategy, results are mixed, but there is still a notable impact. Notably, the results are not directly comparable though due to the usage of different seeds in the split creation.

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 6.9: Stratified Split: Fairness comparison: baseline vs. stratified variants.

Although the stratified variants showed improved fairness, there was a noticeable drop in overall model performance (see Table 6.10). This was somewhat expected, given that the baseline used the full original training set, whereas the stratified variants employed an additional train-validation split, which reduces the number of training samples. Both the F1-score and balanced accuracy decreased compared to the baseline. Strategy B in particular exhibited the poorest performance across most metrics, likely due to its small training set which lacks certain rare cases.

This illustrates the trade-off between fairness and predictive performance that 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 6.10: Stratified Split: Performance comparison: baseline vs. stratified variants.

6.5 Code Contribution

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

7 Outlook

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. Also, it is emphasized to check the Appendix B to get an overview over existing biases and their relevance for PASSION.

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* attribute 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, automated comparison, and image-level analysis.
- 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 used by Valentim et al. (2019).

7.4 Fairness Assessment Results Review

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.

8 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 sociodemographic and hygiene factors influencing impetigo in saudi arabian children: A cross-sectional investigation [Publisher: Dove Medical Press eprint: https://www.tandfonline.com/doi/pdf/10.2147/CCID.S472228]. *Clinical, Cosmetic and Investigational Dermatology*, *17*, 2635–2648. https://doi.org/10.2147/CCID.S472228

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

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.

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.

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.

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

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.

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

Farlex. (n.d.). Pediatric. In *The free dictionary*. Retrieved June 5, 2025, from https://medical-dictionary.thefreedictionary.com/pediatric

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.

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.

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.

Hameed, N., Shabut, A. M., Ghosh, M. K., & Hossain, M. A. (2020). Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. *Expert Systems with Applications*, *141*, 112961. https://doi.org/10.1016/j.eswa.2019.112961

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.

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

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.

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.

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

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.

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

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.

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

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.

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

scikit-learn developers. (n.d.). *StratifiedKFold* [Scikit-learn]. Retrieved June 4, 2025, from https://scikit-learn/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

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.

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, S., Santinelli, M., & Hua, G. (2021, March 11). *Practice AI responsibly with proxy variable detection* [GAMMA — part of BCG x]. Retrieved June 5, 2025, from https://medium.com/bcggamma/practice-ai-responsibly-with-proxy-variable-detection-42c2156ad986

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.

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

Table B.1 summarizes the categorization of bias types. The categories and corresponding biases are described below. Each bias is presented with a definition, an example, a possible mitigation strategy, and its relevance and recommendations for PASSION. Mitigation strategies and examples without citations are based on conclusions drawn from the bias descriptions and the reviewed literature. It is encouraged to further enhance this list.

The italicized part in the following chapter titles indicates the relevance of each category or bias for PASSION (e.g., *high*), based on the following criteria:

- **High.** Directly applicable to PASSION, both in terms of the teledermatology setting and available metadata; likely to provide valuable insights or improvements. Also, crucial demographic biases are included as high, since PASSION's aim is to create more accessible AI models.
- **Medium.** Generally relevant to diagnostic AI and PASSION, but do not seem to have the biggest impact towards PASSION's primary goals.
- **Low.** Related to PASSION but only limited. E.g., in the far future it could potentially impact PASSION.
- **Not Applicable.** Not relevant for PASSION due to fundamental differences in domain, type of data, or type of model. For this categorization, no mitigation methods are listed.

Notably, at the time of writing this thesis, detailed information on the exact data selection process behind PASSION and the rationale for certain decisions was not available. The PASSION relevance sections therefore aim to highlight potential impacts based on the available documentation and observed characteristics of the dataset.

Bias	Mentioned in Context of	
	ML	Dermatology
Data Collection		
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 Biases	X[1]	X[4]
Algorithmic Design		
Algorithmic Biases	X[1]	
External Influence Biases	X[1]	X[4]
User Interactions		
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) [4] (Chakraborty, 2024) [7] (Mester, 2017)
[2] (HP, 2022) [5] (Young et al., 2020) [8] (Delgado-Rodríguez & Llorca, 2004)
[3] (Mester, 2022) [6] (Montoya et al., 2025)

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

B.1 Category: Sampling Bias, *high*

Sampling biases occur when the process of collecting data results in samples that are not representative of the broader population. These biases affect the generalizability of ML models, especially in medical applications, where population diversity is crucial. Non-random or selective sampling can lead to serious consequences in terms of fairness and effectiveness of AI systems (HP, 2022; Mehrabi et al., 2021).

B.1.1 Sampling Bias, *high*

- **Definition:** Bias introduced through non-random sampling of subgroups, leading to poor generalization (Mehrabi et al., 2021).
- **Example:** An ML model trained predominantly on patients with low pigmented skin may underperform on images of patients with highly pigmented skin (Gottfrois et al., 2024).
- **Mitigation Strategy:** Ensure a truly random and inclusive sampling strategy across diverse factors, e.g., by including lots of diverse sources.

- **PASSION Relevance:** PASSION already aims to address sampling bias against highly pigmented skin (Gottfrois et al., 2024). However, if the included data is not truly representative across populations (e.g., over-representation of certain countries, sex, age, skin tone), it could still result in sampling bias. It should be noted that the less pigmented skin tones should also be included in the PASSION dataset to ensure the model generalize among all FSTs.

B.1.2 Selection Bias, *high*

- **Definition:** This bias arises when only a specific subset of the population is used, which is not representative (Chakraborty, 2024; Mester, 2017, 2022).
- **Example:** Training a model only on adult data, when the target population includes children.
- **Mitigation Strategy:** Ensure that data exists for all possible values of metadata attributes and their combinations.
- **PASSION Relevance:** PASSION may be affected by selection bias regarding age, FST, country, and even target labels, as not all possible values are represented. For example, the category "Others" in the condition labels suggests that additional, unlabeled categories exist beyond the three explicitly named. Also, there are more countries in Sub-Saharan Africa - however, the intention of the country attribute should be refined before trying to cover all countries, as the effort would probably be massive. If bias exists for the country, also assess whether it is the symptom of another bias source.

B.1.3 Systematic Selection Bias, *high*

- **Definition:** A form of selection bias where chosen samples differ systematically from the general population (**c5**; **c6**; **c33**; Chakraborty, 2024).
- **Example:** Including only hospitalized patients in a dataset, while most cases are treated in without hospitalization. This especially occures in studies conducted in regional referral centers (**c5**; **c6**; **c33**; Chakraborty, 2024).
- **Mitigation Strategy:** Include mild, moderate, and severe cases from various clinical settings.
- **PASSION Relevance:** If PASSION uses data only from dermatology centers treating severe cases, it introduces systematic selection bias. In that case, other sources should be added. Other possible systematic biases should be checked.

B.1.4 Ascertainment Bias, *high*

- **Definition:** A systematic distortion arising from the method by which participants or data are selected for inclusion (**c5**; Chakraborty, 2024).
- **Example:** Studies on STD prevalence conducted only in public clinics may

overlook patients from higher-income backgrounds who go to private practitioners (**c5**; Chakraborty, 2024).

- **Mitigation Strategy:** Use allocation concealment and blinding to avoid this bias (**c5**; Chakraborty, 2024). Also, ensure that data is collected from a diverse range of sources, e.g., including both public and private healthcare facilities.
- **PASSION Relevance:** If PASSION's dataset is composed mostly of patients from certain types of clinics, it may not generalize well to other socioeconomic groups. PASSION's metadata would need to enhanced with such information to find these kind of impairments. There could be further distortions, therefore, the selection process should be reviewed critically.

B.1.5 Availability Bias, *high*

- **Definition:** Overreliance on easily accessible data rather than the most representative data (**c9**; **c10**; Chakraborty, 2024).
- **Example:** Using only online available datasets for skin conditions may underrepresent rare diseases.
- **Mitigation Strategy:** Actively seek underrepresented data sources, especially for less common or less documented skin types.
- **PASSION Relevance:** PASSION is taking efforts to tackle this bias in dermatology AIs overall by gathering data on FST III to VI. Within the project, PASSION may reflect availability bias by primarily sourcing data from clinics or sources that were most accessible during collection, potentially excluding patients in remote or underrepresented regions within Sub-Saharan Africa. If that is the case, it is crucial that PASSION finds a way to include those regions as well, as they are probably the most vulnerable.

B.1.6 Survivorship Bias, *medium*

- **Definition:** Only using data from so-called survivors, i.e., subjects that make it through a certain threshold or are retained in the dataset, ignoring those who were lost earlier (**Silfwer_2017**; Mester, 2022).
- **Example:** The original example of the World War II airplanes connects the bias indeed to only looking at data of surviving subjects (**Silfwer_2017**). However, as Mester (2022) indicates, also data records which survived preselections in the data collection process can be survivors. E.g., evaluating the success of a treatment based only on patients who completed it, ignoring those who dropped out due to side effects could lead to survivorship bias.
- **Mitigation Strategy:** Account for dropout rates and include cases from a wide range of medical access points.
- **PASSION Relevance:** Patients who are unable to attend the clinics may be excluded from the current dataset, potentially introducing survivorship bias. A more severe implication arises if certain dermatological conditions

are life-threatening or correlate with other lethal diseases—these cases may be underrepresented. Therefore, efforts should be made to ensure such conditions are included in the dataset where possible, especially the early stages to allow for an early triage.

B.2 Category: Representation Biases

Representation biases occur when a sample used to train or evaluate a ML 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 (Mehrabi et al., 2021).

B.2.1 Representation Bias, *high*

- **Definition:** Representation bias arises based on how the sampling during the data collection project is conducted. Non-representative sampling leads to missing subgroups or other representation anomalies, leading to missing or misrepresented characteristics in the data. Popular ML datasets suffer from this bias (Mehrabi et al., 2021; Shankar et al., 2017).
- **Example:** If a skin disease detection model is trained predominantly on FST I-III, it may struggle to accurately diagnose conditions in individuals with darker skin tones (FST IV-VI), as seen in available dermatology models (Gottfrois et al., 2024).
- **Mitigation Strategy:** A potential mitigation strategy could involve ensuring a more balanced representation and periodical reassessment to ensure inclusion over time.
- **PASSION Relevance:** PASSION attempts to mitigate representation bias in existing dermatology AI models 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. It is suggested to include rare and diverse skin conditions and other demographic factors.

B.2.2 Population Bias, *high*

- **Definition:** Population bias occurs when the sample's demographic characteristics (such as age, sex, or ethnicity) do not align with the target population, leading to non-representative data (Mehrabi et al., 2021; Olteanu et al., 2019).
- **Example:** Hargittai (2007) mentioned multiple demographic biases related to this bias, e.g., ethnicity. If a dataset is predominantly comprised of one ethnic group, a dermatology model trained on this data may not generalize well to other ethnic groups, if e.g., the manifestation of skin diseases varies across ethnicities.

- **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.
- **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. In order to evaluate the ethnic diversity, corresponding metadata should be added to the model, if the manifestations of skin diseases indeed vary among ethnicity.

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 sex or ethnicity) are not properly accounted for (Mehrabi et al., 2021; Suresh & Guttag, 2021).
- **Example:** A diagnostic model trained on a heterogeneous dataset might fail to capture how skin diseases manifest differently across sexes or ethnic groups or more dimensional subgroups, potentially leading to misdiagnosis or unequal treatment recommendations (Mehrabi et al., 2021; Suresh & Guttag, 2021).
- **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 (Mehrabi et al., 2021; Suresh & Guttag, 2021). It is especially important to do the analysis based on the model results, as the bias can also occur when (sub-)groups are equally represented in the data (Mehrabi et al., 2021).
- **PASSION Relevance:** Aggregation bias is a significant concern in PASSION, since it involves multiple sensitive demographic factors which impact skin disease prevalence and appearance. The model needs to account for these factors to avoid generalized conclusions that might harm certain subgroups.

B.2.4 Simpson's Paradox, *high*

- **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 (Mehrabi et al., 2021).
- **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.

- **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.
- **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 per FSTs and sexes may appear high, but specific skin conditions in certain FST-sex-subgroups could have lower accuracy when analyzed separately.

B.3 Category: Measurement Biases, *medium*

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 (Mehrabi et al., 2021; Suresh & Guttag, 2021).

B.3.1 Measurement Bias, *high*

- **Definition:** Measurement bias occurs when features or metadata attributes are inaccurately measured or selected (e.g., proxy variables), leading to incorrect interpretations of the outcome (Mehrabi et al., 2021).
- **Example:** A proxy variable it could lead to misinterpretation of the data (S. Wang et al., 2021). For instance, the country of origin or FST may not directly correlate with ethnic or genetic background, potentially skewing results if those factors impact skin conditions.
- **Mitigation Strategy:** To mitigate measurement bias, careful consideration should be given to the choice of attributes used in the dataset. Avoiding proxy variable relevant factors could improve the accuracy of the data and its interpretation.
- **PASSION Relevance:** In the context of the PASSION dataset, measurement bias could arise if country of origin is misused as a proxy variable for ethnicity. The country of origin 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. Instead focus on medically relevant attributes.

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 collec-

tion process, or when different observers report the same observation differently (**c29**; **c26**; Chakraborty, 2024; Mester, 2022).

- **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 (Chakraborty, 2024).
- **Mitigation Strategy:** To address observer bias, standardized training for annotators and a clear, objective set of criteria for diagnosis could be implemented (Montoya et al., 2025). Additionally, using multiple annotators and cross-checking results can help reduce the impact of individual biases.
- **PASSION Relevance:** In PASSION, observer bias could affect the consistency and reliability of skin disease annotations. Different personal experiences could lead to inaccurate classifications, particularly for diseases that are subjective in appearance.

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:** According to Montoya et al. (2025), several factors like the scale order or image context can change how an annotator labels the skin tone, due to personal or cultural biases.
- **Mitigation Strategy:** To reduce annotator bias, *"greater transparency, standardized procedures, and careful consideration of annotator biases"* are needed (Montoya et al., 2025). Maybe, the use of automated tools for initial labeling could provide more objectivity in the process.
- **PASSION Relevance:** In PASSION, annotator bias could particularly affect the labeling of skin tones, which are somewhat dependent on individual perception. This bias could further lead to inconsistent classifications of skin conditions across different demographic groups.

B.3.4 Recall Bias, *low*

- **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 (Chakraborty, 2024; Mester, 2022).
- **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.
- **Mitigation Strategy:** To mitigate recall bias, one could try to gather more objective data through clinical observations or imaging, and ensure

that patient self-reports are validated through medical records or consistent follow-ups.

- **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.

TODO: move down

B.4 Category: Research Biases, *medium*

This category captures the biases which are related to the impacts researcher have on their own studies. Since PASSION has already been published, research biases might already have been introduced. It is not feasible to evaluate this during this thesis. Instead, the PASSION team should check the listed biases and take measures against them if they exist. An external evaluation could help to detect and prevent those biases even further.

B.4.1 Funding / Sponsorship bias

- **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 (**c22**; Chakraborty, 2024; Mester, 2017).
- **Example:** A study funded by a company that produces medications may emphasize the effectiveness of the company's products, even if there is no strong evidence supporting their superiority.
- **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

- **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 (Chakraborty, 2024).
- **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.
- **Mitigation Strategy:** To avoid data dredging, a clear and well-defined hypothesis should be established before conducting any statistical tests. Also,

focusing on confidence intervals and p-curves over p-values could further reduce the bias (Chakraborty, 2024).

B.4.3 Hypothetical bias, *not applicable*

- **Definition:** Hypothetical bias occurs when responses to hypothetical questions do not reflect real-world behavior or preferences (**c31**; **c28**; Chakraborty, 2024).
- **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 PASSION, since PASSION does not involve such kind of questioning.

B.5 Category: Feature Representation Biases, *high*

These types of biases occur when the features or attributes 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 variables are not included in the model, which leads to situations for which the model is not ready for (Clarke, 2005; Mehrabi et al., 2021; Mester, 2017; Mustard, 2003; Riegg, 2008).
- **Example:** A model may accurately predict when customers unsubscribe, but still fails to anticipate a sudden spike in cancellations. A new competitor entered the market with a cheaper alternative. The model did not include this factor (Mehrabi et al., 2021).
- **Mitigation Strategy:** Mitigation can involve incorporating a more comprehensive set of features during training. In medical AI, attributes such as ethnicity or FST may help capture important variations. However, using these variables directly in model training risks introducing new biases. Therefore, it may be preferable to include them in metadata for use in fairness evaluation, rather than as predictive features.
- **PASSION Relevance:** The PASSION dataset has potentially omitted certain sensitive attributes, which could lead to biased results if certain skin diseases and their manifestation vary significantly across different ethnic groups. Without this attributes, the fairness analysis may fail to capture important differences in the data, which could hide omitted variable bias.
  Also, when later an image of healthy skin is uploaded, the model might not be prepared for it. Such data should be included.

B.5.2 Collider Bias, *medium*

- **Definition:** 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).
- **Example:** For example, psoriasis and depression may appear linked because severe cases are hospitalized, where mental health screening occurs. This can create a false association due to collider bias, as both conditions influence the likelihood of hospitalization but are not necessarily linked to each other (Chakraborty, 2024).
- **Mitigation Strategy:** Probably, assess cause and effects and be mindful what data to include in the dataset.
- **PASSION Relevance:** Due to lacking dermatology knowledge, the real impact of this model is currently unclear. The PASSION team should verify whether collider variables exists among the supported dermatological conditions and other factors in the image.

B.6 Category: Imaging Biases, *high*

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 ML models (Young et al., 2020).

B.6.1 Image Quality Bias, *high*

- **Definition:** Image quality bias occurs when the quality of an image, such as the zoom level, focus, lighting, or even different hardware affects how a ML model classifies or diagnoses the image. Poor image quality can lead to misclassification or lower prediction accuracy (Young et al., 2020).
- **Example:** If a dermatologist captures an image with insufficient lighting or poor focus, the model may struggle to identify skin conditions, potentially leading to a misdiagnosis (Young et al., 2020).
- **Mitigation Strategy:** Often, poor-quality images are discarded. However, it would be better, if the model would become more robust. Instead of removing those images from the dataset, define what is an adequate image and let the model assess image quality. If the model can express confidence based on it, it could prompt the users to retake photos if necessary (Young et al., 2020).
- **PASSION Relevance:** Since PASSION will be used in a teledermatology context, it will not be feasible to fully standardize image acquisition which was also proposed by Young et al. (2020). The image quality assessment proposed above is probably the best method going forward.

Also, the biased outcome regarding the countries could be an indicator, that this bias indeed exists in PASSION.

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 (**Winkler_2019**; **Bisla_2019**; Young et al., 2020).
- **Example:** A photograph of a skin lesion may contain hair or markings from previous medical procedures, making it more difficult for the model to identify the skin condition correctly (Young et al., 2020).
- **Mitigation Strategy:** To reduce visual artifact bias, it is important to implement pre-processing steps that remove or mask artifacts in images. This could involve techniques such as hair removal (**Bisla_2019**). Depending on the use case, this could be done algorithmically or even before taking the picture.
- **PASSION Relevance:** Similar to the bias before, this bias is highly relevant and the suggested method should be assessed. Again, due to the teledermatology setup, the hair removal would probably need to be done algorithmically since one can not expect people to shave themselves before uploading images.

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 (**Mishra_2019**; Young et al., 2020).
- **Example:** If only a small portion of a skin lesion is captured in the image or a big portion of healthy skin is captured around, the model may miss critical features needed to correctly identify conditions.
- **Mitigation Strategy:** To address field of view bias, the dataset should ensure that images are captured from standardized and consistent angles or distances (Young et al., 2020). Augmenting the dataset with a variety of views from multiple angles could potentially also help improve the model's ability to generalize to unseen cases.
- **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. Again, the model should be ready for this due to its teledermatology use case. There is already data captured on what body parts are affected by the labeled condition. If this data could be

extended to be captured per image, potentially, more conclusions could be drawn in regards of this bias.

B.7 Category: Medical Biases Originating in Data Collection, *high*

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

B.7.1 Berkesonian Bias, *medium*

- **Definition:** Berkesonian bias occurs in hospital-based studies when two factors (such as disease severity or risk factors) influence independently whether patients seek treatment or are hospitalized. This can distort the relationship between variables due to the study population of hospitalized patients is not representative of the general population (**c3**; **c7**; Chakraborty, 2024).
- **Example:** If a study looks at how pregnancy affects syphilis in an antenatal clinic, the data might be biased because both pregnancy and syphilis influence who attends the clinic and, therefore, the observations (**c3**; **c7**; Chakraborty, 2024).
- **Mitigation Strategy:** To mitigate Berkesonian bias, one could include a diverse set of patients from multiple sources, including both hospital and non-hospital populations, ensuring a more representative dataset.
- **PASSION Relevance:** This bias should be checked in more details. Potentially, the relationship between the target labels conditions and impetigo could be influenced by those biases, leading the model to learn a connection between them.

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 (**c27**; **c23**; Chakraborty, 2024).
- **Example:** A person who is already being treated for one skin condition might also be screened for other conditions, leading to a misinterpretation of relationships between conditions.
- **Mitigation Strategy:** To reduce informed presence bias, the model should probably account for patients with varying levels of care-seeking behavior and ensure that both treated and untreated conditions are represented in the dataset.
- **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. This in turn could influence the predictions for a condition together with impetigo.

B.7.3 Diagnostic Access Bias, *high*

- **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 (Chakraborty, 2024).
- **Example:** The prevalence for atopic dermatitis is believed to be higher in the West than in India, what could be linked to better accessible diagnostic facilities (Chakraborty, 2024).
- **Mitigation Strategy:** To address diagnostic access bias, it one should ensure that the dataset includes a diverse range of geographical locations and healthcare access levels, including both early and late-stage conditions.
- **PASSION Relevance:** PASSION addresses diagnostic access bias in dermatology AIs regarding Sub-Saharan Africa. However, the bias could still be relevant in the dataset, depending on which clinics were chosen for data selection.

B.7.4 Diagnostic Reference Test Bias, *medium*

- **Definition:** This is a *verification bias* which occurs when not all individuals in a study receive the same reference test, leading to discrepancies in diagnoses (Chakraborty, 2024).
- **Example:** When not all patients are diagnosed using the same tests (e.g., a skin biopsy based diagnosis vs. a more thorough procedure) for the same condition, it causes inconsistency in diagnostic results.
- **Mitigation Strategy:** The diagnostic processes should be standardized and applied for all patients in the same way across different healthcare settings. Consistent use of reference tests when collecting data must be ensured.
- **PASSION Relevance:** Depending on how dermatologists work in the PASSION context, diagnostic reference test bias could be present, if different diagnostic methods or reference tests are used while labeling the data,

B.7.5 Mimicry Bias, *medium*

- **Definition:** Mimicry bias occurs when treatment exposure causes a disease that closely resembles the study disease, potentially leading to misleading data (Chakraborty, 2024).
- **Example:** Certain drugs can induce a disease-like reaction, which looks similar to the initial disease but is clinically different (Chakraborty, 2024).

- **Mitigation Strategy:** Careful documentation of treatment histories and known mimicking conditions is essential. Inclusion of additional clinical metadata can help disambiguate mimicked conditions.
- **PASSION Relevance:** Diseases that visually resemble others, could be mistakenly labeled in PASSION if treatment history is not considered. This can negatively affect model accuracy.

B.8 Category: Temporal Biases, *not applicable*

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.

- **Longitudinal Data Fallacy** (Mehrabi et al., 2021)
- **Chronological Bias** (Chakraborty, 2024)
- **Immortal Time Bias** (Chakraborty, 2024)

B.9 Category: Algorithmic Biases, *low*

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, *low*

- **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). *Presentation bias* and *Ranking bias* are further subtypes mentioned by Lerman and Hogg (2014) and Mehrabi et al. (2021).
- **Example:** For instance, if a teledermatology app visually emphasizes certain information, users may begin to prioritize this information, which could distort the results the algorithm provides.
- **Mitigation Strategy:** Careful evaluate the UI design, potentially with a UX designer, ensuring that no unintended prioritization occurs.
- **PASSION Relevance:** User interaction biases could emerge as the PASSION teledermatology platform becomes publicly available. When the en-

tered data would be used for further training, the bias would need to be investigated.

B.9.2 Emergent Bias, *low*

- **Definition:** When real users interact with an algorithm, this bias arises some time after the design was completed due to changes in population. It appears mostly in user interfaces (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.
- **Mitigation Strategy:** Continuous monitoring of how the system interacts with different demographic groups could mitigate this bias. Ensuring that new data from diverse populations is incorporated into the training set periodically can help counteract emergent biases.
- **PASSION Relevance:** Again, this bias is relevant only later to PASSION, when the system will be opened for a 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.
  Even though this bias is only relevant in the future, this is a reason why PASSION's data should also include low pigmented skin types and healthy skin examples.

B.10 Category: External Influence Biases, *high*

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 (Mehrabi et al., 2021). More examples can be found in the work of Young et al. (2020).

B.10.1 Evaluation Bias, *high*

- **Definition:** When inappropriate or disproportionate benchmarks are used in model evaluation, the benchmarks' biases can be introduced into the model (Buolamwini & Gebru, 2018; Suresh & Guttag, 2021).
- **Example:** The *Adience* and *IJB-A* benchmarks were identified as inappropriate benchmarks (Mehrabi et al., 2021).
- **Mitigation Strategy:** Benchmarks should be carefully assessed before using them.
- **PASSION Relevance:** PASSION aims to become a benchmark for dermatology models TODO: cite mid term. Therefore this bias is relevant for

PASSION in terms of being extensive in the fairness assessment to become an appropriate benchmark.

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).
- **Example:** If a model 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.
- **Mitigation Strategy:** Ensuring that the reference tests used for validation are distinct and independent from the model's diagnostic tests can mitigate incorporation bias.
- **PASSION Relevance:** Incorporation bias is probably less relevant for PASSION since it likely relies on independent diagnostic practices to validate its dermatological models, reducing the chance of this type of bias affecting its evaluations.

B.11 Category: Cognitive Biases, *high*

Biases which are related to human perception belong to the category of cognitive biases. These biases can impact how data is presented and interpreted (Mester, 2017).

B.11.1 Confirmation Bias, *medium*

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

B.11.2 Belief Bias, *medium*

- **Definition:** This is essentially a stronger version of the confirmation bias, where judgments are influenced by pre-existing beliefs or intuitions, leading 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).
- **Mitigation Strategy:** Implementing blind labeling processes, where experts are unaware of previous diagnoses, could help reduce belief bias. Also, let multiple experts label the data independently could further mitigate the bias .
- **PASSION Relevance:** This bias could be incorporated in inaccurate diagnosis and labeling if experts rely too heavily on their subjective interpretation of the data rather than objectively evaluating it.

B.11.3 Previous Opinion Bias, *high*

- **Definition:** When the knowledge of prior results or diagnoses influences the interpretation of new data, leading to biased conclusions (Chakraborty, 2024).
- **Example:** A dermatology expert may reach different conclusions depending on whether they know a previous diagnosis beforehand or assess the case without that knowledge.
- **Mitigation Strategy:** Labeling experts should independently diagnose cases without access to previous diagnoses.
- **PASSION Relevance:** This is not only relevant to PASSION's labeling process but even more importantly, it also affect real-world diagnoses once PASSION is deployed. For example, if both the patient and the dermatologist are aware of the model's prediction before the dermatologist evaluates the case, the model's output could influence the final diagnosis. Therefore, it is crucial that the model's prediction is not shown to users - at least not during triage and prior to the clinical assessment - unless it concerns a condition that users can safely treat themselves.

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:** Children who were tutored probably got worse grades than other children. However, the bad graded caused the tutoring, not vice-versa (Mester, 2017).

- **Mitigation Strategy:** According to Mester (2017), the only way to assess cause-effect is via experimenting.
- **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.

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 typically privileged subgroup of the population may not accurately represent skin diseases in other populations (Mehrabi et al., 2021).
- **Mitigation Strategy:** Ensuring diversity in the dataset by collecting data from a wide range of demographic groups (age, sex, race, etc.) is essential to reduce historical bias. Efforts should be made to balance the dataset and account for historically marginalized groups.
- **PASSION Relevance:** PASSION addresses this bias by providing a dermatology dataset of usually underrepresented FSTs. However, it should also be noted that the Fitzpatrick scale is historically skewed towards low pigmented skin types (Montoya et al., 2025). Therefore, other scales should be assessed to get a more accurate labeling of skin types to allow for a more complete fairness assessment.

B.11.6 Content Production Bias, *high*

- **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 in a teledermatology setup, leading to a potential misrepresentation of the condition's appearance in the data of users.
- **Mitigation Strategy:** The same methods apply as for the image quality bias subsection B.6.1.
- **PASSION Relevance:** The same relevance applies as for the image quality bias subsection B.6.1. TODO: check appearance in document of references in document

B.12 Category: Behavioral Biases, *high*

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:** User behavior can differ depending on 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.
- **Mitigation Strategy:** A diverse set of data from various geographical and cultural backgrounds should be included in the training dataset to overcome the bias.
- **PASSION Relevance:** Differences in healthcare-seeking behavior across cultures or countries may lead to an unrepresentative sample in PASSION's dataset. Therefore, including data from various countries could help account for these differences and improve the generalizability of the model. PASSION already covers 4 countries, but the balance could be improved. Also, it should be considered to add more countries.

B.12.2 Self-Selection Bias, *high*

- **Definition:** 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 (Mehrabi et al., 2021; Mester, 2022). *Non-responder bias* is a subtype, where part of the population is not responding e.g., to fill out a survey or post-study questionnaires (Chakraborty, 2024).
- **Example:** When only patients who seek dermatological care at hospitals would be included, which could exclude individuals with skin conditions who do not seek medical help.
- **Mitigation Strategy:** Trying to gather information via other data sources.
- **PASSION Relevance:** Self-selection bias is a significant issue for PASSION since the dataset relies on patients who visit clinics, meaning those who do not seek treatment or who do not have access to healthcare will be underrepresented in the dataset. Maybe, there are organizations which are promoting more proactively treatment options to potential patients. They could be included in the data gathering.

B.12.3 Social Bias, *low*

- **Definition:** When the actions of others affect our judgment (Baeza-Yates, 2018).
- **Example:** Ratings in juries can be affected by this bias (Baeza-Yates, 2018).
- **PASSION Relevance:** Since PASSION does not build on social bias, the bias seem irrelevant. Unless social bias could affect whether or not patients seek treatment for a condition depending on what their peers do. This could again lead to data skews, since some conditions would get treated more often than others.

B.13 Category: Publication Biases, *text*

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. To mitigate them, publishers should be open-minded for alternative explanations (Chakraborty, 2024).

B.13.1 Hot Stuff Bias, *medium*

- **Definition:** Hot stuff bias refers to the tendency for journals to be less critical of research related to trending topics, leading to the disproportionate publication of these studies (Chakraborty, 2024).
- **Example:** During the COVID-19 pandemic, numerous journals published studies on cutaneous manifestations of the virus (Chakraborty, 2024).
- **Mitigation Strategy:** Reviewers can reduce this bias by applying consistent publication standards and avoiding the temptation to prioritize overly trending research without sufficient evidence (Chakraborty, 2024).
- **PASSION Relevance:** In today's world, fairness across sex, skin types, and ethnicities is a widely discussed topic and these factors should certainly be considered in PASSION. However, it is crucial to evaluate their actual impact on dermatological conditions. Focusing solely on popular subgroup categories does not help if these factors have no meaningful connection to the diseases being studied. And other relevant variables must also be taken into account to ensure a truly fair and effective evaluation.

B.13.2 All is Well Bias, *high*

- **Definition:** This bias occurs when theories that align with the majority or dominant views are more likely to be published than those that challenge the consensus (Chakraborty, 2024).
- **Example:** Theories supporting autoimmunity as the cause of endemic pemphigus are more likely to be published than those suggesting an infectious origin, favoring certain viewpoints (Chakraborty, 2024).

- **Mitigation Strategy:** This bias seem to be *"very difficult to eliminate"* (Chakraborty, 2024). Probably, as a first step in the right direction, certain parties need to make greater efforts to pave the way for others.
- **PASSION Relevance:** In a way, PASSION took the first step to include data on highly pigmented skin types which seem to be harder to collect, because they recognized that there seem to be a disparity in the current research (Gottfrois et al., 2024).
  However, the Fitzpatrick scale goes into that direction as well. New research, e.g., by Montoya et al. (2025) suggest the long standing standard in dermatology is flawed. The scale and alternatives should be assessed.

B.13.3 Rhetoric Bias, *medium*

- **Definition:** Charismatic writing or when the press is more vocal about findings can lead to greater influence over individuals than other available facts (Chakraborty, 2024).
- **Example:** The wider use of sunscreen over other protective measures like umbrellas or hats for sunlight may be due to stronger promotion of sunscreens in the press (Chakraborty, 2024).
- **Mitigation Strategy:** Researchers should be aware of this bias while reading and stick with neutral wording when writing papers. Journals and reviewers should also ensure that rhetoric does not overshadow the actual scientific contribution of a paper.
- **PASSION Relevance:** This bias is as relevant for PASSION as for all research projects.

B.13.4 Novelty Bias, *high*

- **Definition:** Newer interventions appear to be better, even if the evidence does not support this. Over time, this effect decreases (Chakraborty, 2024).
- **Example:** A new medicine may be reported as highly effective. However, independent, larger studies show less impressive outcomes.
- **Mitigation Strategy:** New approaches should be compared to established methods in controlled studies. Reviewers should ensure that novelty does not overshadow the importance of replicability and robustness in research findings.
- **PASSION Relevance:** The previously mentioned assessment of the Fitzpatrick scale by (Montoya et al., 2025) seem to be new. Further studies investigating the scale should be reviewed (or conducted) to get a more precise insight on the effects.

B.14 Category: Medical Biases Originating in User Interactions, *high*

Again, medical user interaction biases are more healthcare-specific biases. The difference to the previous chapter is, that this

B.14.1 Popularity Bias, *high*

- **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). Additionally, popularity bias occurs when more well-known or stigmatized diseases are overrepresented in healthcare settings compared to less common diseases. This can result in a distorted view of the prevalence and severity of different conditions (Chakraborty, 2024).
- **Example:** If the training data focuses too heavily on commonly encountered dermatological conditions or frequently observed features, the model may fail to correctly diagnose rarer skin diseases due to their underrepresentation.
  This can be amplified by patients reaction to widely recognized or stigmatized conditions, resulting in the overrepresentation of such diseases while rarer disorders receive less attention (Chakraborty, 2024).
- **Mitigation Strategy:** To mitigate popularity bias, it is important to ensure that the training dataset includes both common and rare skin conditions, offering a comprehensive representation of dermatological diseases.
  This could involve actively seeking data from hospitals or clinics that treat a wider variety of dermatological conditions, including rare ones. Specialists on rare conditions could also be considered to supplement a dataset.
- **PASSION Relevance:** Since PASSION focuses on the most common skin conditions according to Gottfrois et al. (2024), this bias is particularly relevant. While it makes sense to tackle the most common conditions first to have a big impact, it should not be forgotten to include rarer skin conditions to be inclusive.

B.14.2 Apprehension Bias, *not applicable*

- **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 (Chakraborty, 2024).
- **Example:** A patient may have elevated blood pressure readings due to anxiety before a dermatological procedure, leading to an inaccurate diagnosis or assessment (Chakraborty, 2024).
- **PASSION Relevance:** This bias is considered irrelevant for PASSION, as the likelihood of fear influencing the captured images is minimal. Nonethe-

less, it remains important to ensure that patients feel comfortable with the imaging process.

B.14.3 Hawthorne Bias, *medium*

- **Definition:** Subjects might modify their behavior when they know they are being watched (Chakraborty, 2024).
- **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.
- **Mitigation Strategy:** This bias could only be mitigated by minimizing participants' awareness of being observed, which, however, raises ethical concerns and is not recommended.
- **PASSION Relevance:** This bias can be practically utilized as indicated by Chakraborty (2024). In PASSION, it could be used by introducing control mechanisms in the process where professional assessment is relevant.

B.14.4 Centripetal Bias, *medium*

- **Definition:** 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 (Chakraborty, 2024).
- **Example:** Well-known cosmetologists with strong reputations are more likely to attract a more cases compared to their lesser-known peers (Chakraborty, 2024).
- **Mitigation Strategy:** Ensuring diversity in regards of data sources by including data from both specialized and general clinics.
- **PASSION Relevance:** PASSION's choice of partner clinics may introduce this bias, which should be revised.
  However, as indicated in subsection B.14.1, the effects of this bias could be leveraged to include rare conditions by collaborating with a specialist focused on those cases.

B.14.5 Unacceptable Disease Bias, *medium*

- **Definition:** This bias arises when diseases that are socially stigmatized or culturally sensitive are underreported (Chakraborty, 2024).
- **Example:** Patients may avoid seeking medical help for conditions like leprosy diseases (Chakraborty, 2024).
- **Mitigation Strategy:** Awareness campaigns, anonymized data collection, and improved access to care could potentially reduce underreporting.
- **PASSION Relevance:** PASSION could be affected.

B.14.6 Healthy Volunteer Selection Bias, *low*

- **Definition:** A form of self-selection bias where individuals who volunteer for studies tend to be healthier or more health-conscious than the general population (Delgado-Rodríguez & Llorca, 2004; Mehrabi et al., 2021).
- **Example:** The UK Biobank study found that participants were generally healthier than the population, potentially biasing research findings (Mehrabi et al., 2021).
- **Mitigation Strategy:** Recruitment efforts should target diverse populations, including those with poor health or low access to care.
- **PASSION Relevance:** PASSION is probably not affected by this bias, since data were collected in clinical settings. However, the data gathering processes could be reviewed, as the bias might also manifest in the opposite direction.

C Fairness Metrics

According to Mehrabi et al. (2021), fairness can be achieved at group, subgroup or individual level. Group fairness involves treating different groups equally. Individual fairness aimes to achieve similar predictions for similar individuals. Subgroup fairness incorporates 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 a list of fairness definitions, structured according to these 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 aim to achieve similar probability outcomes for 'unprotected' and 'protected' groups. The following list summarizes how they work:

- **Demographic/Statistical Parity** and **Conditional Statistical Parity**: The parity checks that the likelihood of a positive outcome is equal 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 if they belong to the positive class. These methods essentially compare the groups' TPRs. **Equalized Odds** is more restrictive as it also checks for similar false positive rates (Mehrabi et al., 2021; Verma & Rubin, 2018).

- **Treatment Equality**: This method compares the false negative and false positive rates (T. Wang & Wang, 2014)
- **Counterfactual Fairness**: This approach differs from the others in that it tests the same individual in both different demographic groups with the intention of achieving the same outcome (Kusner et al., 2017; Mehrabi et al., 2021). Unlike the first group of fairness metrics, it does not compare the likelihoods of outcomes for any individual within groups, rather, it checks how the exact same individual would be treated if they were in another group.
- **Fairness Through Awareness**: This method compares similar individuals based on similarity metrics to achieve a similar outcome (Dwork et al., 2012; Mehrabi et al., 2021)
- **Fairness Through Unawareness**: This measure ensures that protected attributes are not used explicitly used in decision-making (Grgic-Hlača et al., 2016; Kusner et al., 2017).
- **Fairness in Relational Domains**: This notion also considers relational structures between individuals (Farnadi et al., 2018).

D List of Mitigation Methods

An overview of existing mitigation methods is shown below. It is divided into three tables: Table D.1 presents fairness-oriented data practices, Table D.2 covers fair classification methods, both relevant for PASSION. For completeness, Table D.3 lists selected methods for other ML tasks.

Mitigation Methods - Fair Data Collection and Design	**Context**	
	ML	**Dermatology**
Documentation and Transparency		
Good practices while using data	X[1,2,3]	
Datasheets for dataset creation method, characteristics, motivations and skews	X[1,2,3]	
Datasheets for model method, characteristics, motivations and skews	X[1,4]	
Dataset (nutrition) labels	X[1,5,6]	X[17]
Publish datasets accessible for the public		X[17]
Bias Detection and Evaluation		
Test for Simpson's Paradox	X[1,7,8,9]	
Detect direct discrimination with causal models	X[1,10]	
Out-of-distribution detection in dermatology using input perturbation and subset scanning		X[18]
Confidence intervals and p-curve over p-values		X[16]
Study Design		
Allocation concealment and blinding		X[16]
Preventing direct and indirect discrimination	X[1,11]	
Stratified Splitting	X[19]	
Data Gathering		
Data collection from diverse sources	X[17]	
Robust standards for external validation	X[17]	
Preferential sampling	X[1,12,13]	
Geographical diversity in dataset creation	X[15]	
Balanced skin tone and gender representation		X[18]
Disparate impact removal	X[1,14]	
Labeling		
Multidimensional scale for skin tones		X[18]

[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] (**M75_**)
[13] (**M76_**)
[14] (**M51_**)
[15] (Shankar et al., 2017)
[16] (Chakraborty, 2024)
[17] (Young et al., 2020)
[18] (Montoya et al., 2025)
[19] (F. Chen et al., 2024)

Table D.1: Mitigation Methods Overview: Fair Data Collection and Design

Mitigation Methods - Fair Classification	Mentioned in Context of ML	Dermatology
Satisfy Fairness Definitions		
Satisfy Subgroup Fairness	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]	
Other Fairness Metrics	X[1,7, 8, 9, 10]	
Satisfy Fairness and Stability Under Distribution Shifts	X[1,11]	
Fair Representation Learning		
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]
Others		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 D.2: Mitigation Methods Overview: Fair Classification

Mitigation Methods - For Other ML Tasks	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]	
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 Methods	X[1]	
Causal Fairness and Disparate Learning		
Disparate Learning Processes (DLP)	X[1,9]	
Causal Approach to Fairness	X[1]	
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 D.3: Mitigation Methods Overview: For Other ML Tasks

E PASSION Dataset Distribution Analysis

The data in Table E.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 E.1, the data is visualized.

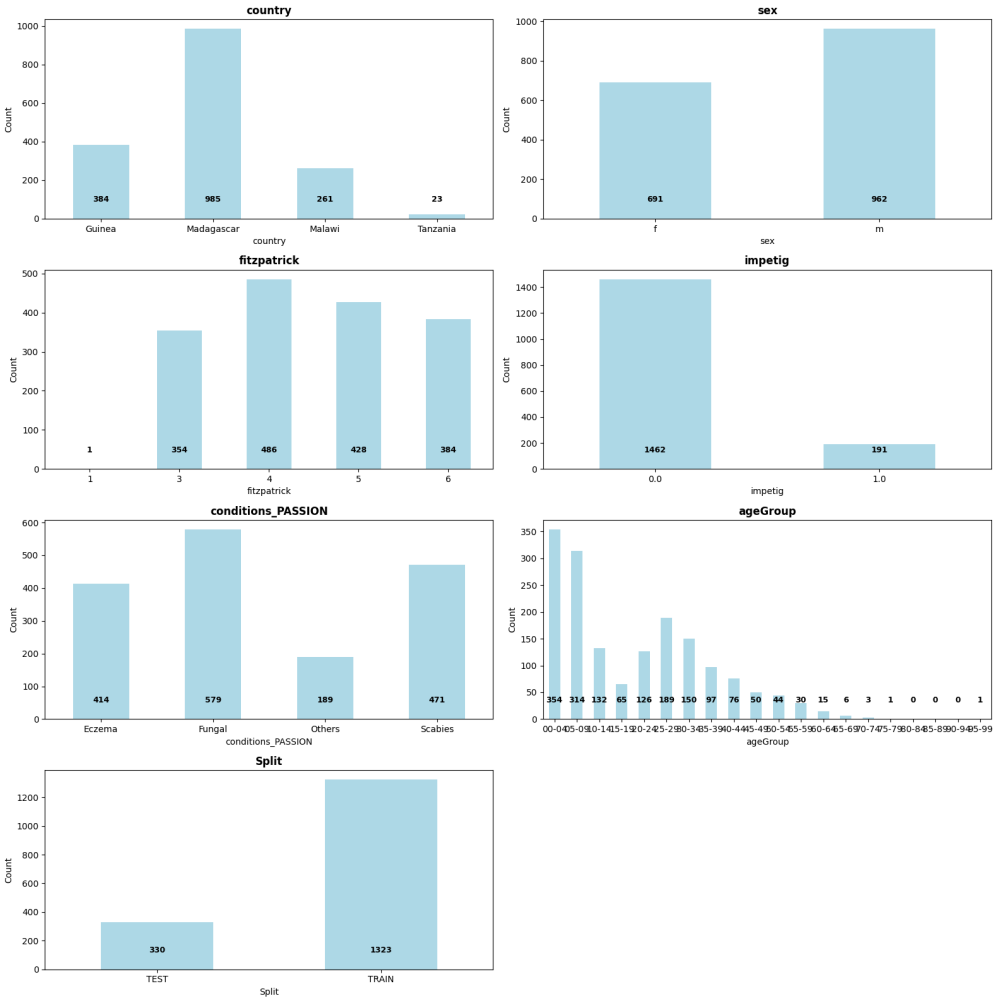

Figure E.1: PASSION dataset distribution analysis on group level

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 E.1: Distribution of metadata attributes in the PASSION dataset

TODO: check the gls all unused.