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

Demographic Biases in Dermatology Models

TODO: subtitle

Bachelor Thesis

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

by

Nadja Stadelmann

from

Lucerne (Switzerland)

Declaration

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

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

Code/Thesis Classification

☒ Public (Standard)
☐ Private

Declaration

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

Place/Date, Signature _____

Submission of the Thesis to the Portfolio Database

Confirmation by the student

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

Place/Date, Signature _____

Expression of thanks and gratitute

Thanks to my family, relatives and friends for all the support given to finish this thesis. TODO: add thanks and gratitude Ludovic Amruthalingam Simone Lionetti - deputy Ludovic Pascal Baumann - LaTeX Philippe Gottfrois - information and work on PASSION project

Nadja Stadelmann, 2025

Summary

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

Contents

Todo list

TODO: solve todos

TODO: also solve todos in the code ;)

TODO: also fix metadata entry!!!

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

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

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

TODO: Portfolio DB für Referenzarbeiten anschauen

List of Figures

List of Tables

Glossary

Fitzpatrick skin type A skin classifier based on the skins' reaction to light (Gottfrois et al., 2024). ix, 2,

Jupyter Notebook Executable files, often used in ML to write Python code and add explanations in text form. 4

pediatric A medical term for infants, children and adolescents. 1, 2

Acronyms

FST Fitzpatrick skin type. *Glossary:* Fitzpatrick skin type, 2, 4, 5, 29

HSLU long. 2

1 Problem Statement

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

TODO: Formulate statement from those citations:

- AI systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations (Mehrabi et al., 2021).
- There are clear benefits to algorithmic decision-making; unlike people, machines do not become tired or bored [45, 119], and can take into account orders of magnitude more factors than people can. However, like people, algorithms are vulnerable to biases that render their decisions "unfair" [6, 121]. In the context of decision-making, fairness is *the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics*. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people. (Mehrabi et al., 2021).
- it is important for researchers and engineers to be concerned about the downstream applications and their potential harmful effects when modeling an algorithm or a system (Mehrabi et al., 2021).
- We should think responsibly, and recognize that the application of these tools, and their subsequent decisions affect peoples' lives; therefore, considering fairness constraints is a crucial task while designing and engineering these types of sensitive tools (Mehrabi et al., 2021).

1.1 Context

This thesis is part of the PASSION project. The PASSION research team identified that in Africa, dermatology treatment is not accessible. There is less than one dermatologist per one million citizens. In contrast, there is high demand for dermatology treatment, especially among children and adolescents. 80% of the pediatric population is affected. The goal of PASSION is to make dermatology treatment more accessible by using AI supported telemedicine for triage (Gottfrois et al., 2024).

For AI supported triage, demographic biases in existing dermatology models is a problem since the corresponding datasets lack diversity, especially regarding skin tones (Gottfrois et al., 2024). This type of bias is important in dermatology, since different diseases present themselves differently depending on the skin-color (Diaz et al., 2022). Further, skin diseases are more advanced or severe at diagnosis in patients with lower socioeconomic status (British Association of Dermatologists (BAD), 2021).

PASSION tries to mitigate the demographic bias by providing a dataset of pigmented skin images of patients from Sub-Saharan Africa. The PASSION team focused on gathering data with Fitzpatrick skin type (FST) IV, V and VI. Further, the covered conditions represent up to 80% of the conditions in the pediatric population, the demographic group who is most affected by skin disease (Gottfrois et al., 2024). long (HSLU)

The PASSION dataset is complementary to the existing datasets and improves the diversity in a combined dataset. Within the dataset itself, there could potentially be further demographic biases, e.g. related to age or gender.

1.2 Objective

The goal of this research is to

1. Identify demographic biases in dermatology AI models, using established fairness metrics.
2. Identify mitigation strategies to minimize these biases.
3. Assess the effectiveness of the mitigation strategies.

It is important to identify the existent biases first, so that the mitigation strategies can be TODO: proceed here to reason why you chose those objectives

2 State of Research

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

2.1 PASSION for Dermatology

The PASSION research team provides a dataset including three analysis scripts and an AI model. For this thesis, it is important to understand which labels the dataset provides, so that the applicable bias mitigation methodologies can be chosen.

The provided analysis scripts show a first insight into the demographic distribution in the dataset, such as Fitzpatrick skin type and cases per country distribution. The results of those analyses reveal first biases.

There are also dermatology specific analysis scripts in regards of body localization by condition or impetigo cases. Those results

2.1.1 PASSION Dataset

The PASSION dataset contains data from patients from four African countries. It contains 4901 images of dermatology cases and the corresponding demographic and clinical information, see Table 2.1. There is one record per patient and one or more corresponding images. The images are linked with the record by filename, which contains the subject_id of the row entry. Access to the dataset can be requested via https://passionderm.github.io/ (Gottfrois et al., 2024).

Label	Data Type	Description
subject_id	string	Participant's unique identifier
country	string	Country of origin of the participant
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)	Specific 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 - labels and descriptions (Gottfrois et al., 2024)

2.1.2 PASSION Analysis Scripts

With the Dataset, the PASSION research team provides a Jupyter Notebook with code examples and analysis scripts. They are listed in Table 2.2 with a description and an indicator, how relevant the scripts are for this thesis.

Script Title	Description	Relevance - Reasoning
Linking CSV Data with Image Files	Creates mapping between the data records and images. It further counts the cases by country	**High** - Basis for other analysis's, potentially provides dermatological info
Extracting and Comparing Subject IDs	Checks the dataset complecity and accurracy in regards of linking records and images	**Low** - Checks loaded data for completeness, but is not providing more insight
Regrouping Malawi and Tanzania to EAS	data aggregation due to dataset size and geographical proximity	**Low** - Might be relevant to understand the dataset and for interpreting the results of the following scripts correctly
Conditions by Country	Relationship between clinical conditions and country	**Medium** - Currently unsure whether this information is relevant for this thesis TODO: research relevance between country vs. clinical conditions in regards of demographic bias
Body Localizations by Conditions	Shows correlation between the condition and primarily affected body parts; does not use all affected body parts listed in the data TODO: check with Philippe why this was done	**Low** - While the correlation can be interesting for other research, it is not relevant for demographic biases.
Impetigo Cases	Counts total number of impetigo cases as well as proportion to all cases	**Medium** - Currently unsure whether this information is relevant for this thesis TODO: research relevance between impedigo and demographic bias
Distribution of Fitzpatrick Skin Types	Counts and visualizes the skin type distribution	**High** - FST is a demographic information

Table 2.2: PASSION dataset - existing analysis scripts (Gottfrois et al., 2024) TODO: decide on a table style

2.1.3 PASSION Experiments

see https://github.com/Digital-Dermatology/PASSION-Evaluation

2.2 General ML biases

2.2.1 Discrimination vs. Biases

- bias and discrimination = source of unfairness. Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally, while bias can be considered as a source for unfairness that is due to the data collection, sampling, and measurement. Although bias can also be seen as a source of unfairness that is due to human prejudice and stereotyping, in the algorithmic fairness literature it is more intuitive to categorize them as such according to the existing research in these areas. In this survey, we mainly focus on concepts that are relevant to algorithmic fairness issues. (Mehrabi et al., 2021)

- Explainable Discrimination. Differences in treatment and outcomes amongst different groups can be justified and explained via some attributes in some cases. In situations where these differences are justified and explained, it is not considered to be illegal discrimination and hence called explainable [77]. In [77], authors present a methodology to quantify the explainable and illegal discrimination in data. They argue that methods that do not take the explainable part of the discrimination into account may result in non-desirable outcomes, so they introduce a reverse discrimination which is equally harmful and undesirable. They explain how to quantify and measure discrimination in data or a classifier's decisions which directly considers illegal and explainable discrimination.(Mehrabi et al., 2021)

- Unexplainable Discrimination. In contrast to explainable discrimination, there is unexplainable discrimination in which the discrimination toward a group is unjustified and therefore considered illegal. Authors in [77] also present local techniques for removing only the illegal or unexplainable discrimination, allowing only for explainable differences in decisions. These are preprocessing techniques that change the training data such that it contains no unexplainable discrimination. We expect classifiers trained on this preprocessed data to not capture illegal or unexplainable discrimination. Unexplainable discrimination consists of direct and indirect discrimination.(Mehrabi et al., 2021)

  - Direct Discrimination. Direct discrimination happens when protected attributes of individuals explicitly result in non-favorable outcomes toward them [164]. ... these traits that are considered to be "protected" or "sensitive" attributes in computer science literature (Mehrabi et al., 2021)
  - Indirect Discrimination. In indirect discrimination, individuals appear to be treated based on seemingly neutral and non-protected attributes; however, protected groups, or individuals still get to be treated unjustly

as a result of implicit effects from their protected attributes (Mehrabi et al., 2021)

- Discrimination can be either direct or indirect. Direct discrimination occurs when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made based on nonsensitive attributes which are strongly correlated with biased sensitive ones (**M62_**)

2.2.2 Bias Introduction

- These biased predictions stem from the hidden or neglected biases in data or algorithms (Mehrabi et al., 2021).
- two potential sources of unfairness in machine learning outcomes - those that arise from biases in the data and those that arise from the algorithms ... we observe that biased algorithmic outcomes might impact user experience, thus generating a feedback loop between data, algorithms and users that can perpetuate and even amplify existing sources of bias (Mehrabi et al., 2021).
- Bias in facial recognition systems [128] and recommender systems [140] have also been largely studied and evaluated and in many cases shown to be discriminative towards certain populations and subgroups. In order to be able to address the bias issue in these applications, it is important for us to know where these biases are coming from and what we can do to prevent them.(Mehrabi et al., 2021).
- We should think responsibly, and recognize that the application of these tools, and their subsequent decisions affect peoples' lives; therefore, considering fairness constraints is a crucial task while designing and engineering these types of sensitive tools (Mehrabi et al., 2021).
- compared SAVRY, a tool used in risk assessment frameworks that includes human intervention in its process, with automatic machine learning methods in order to see which one is more accurate and more fair. Conducting these types of studies should be done more frequently, but prior to releasing the tools in order to avoid doing harm (Mehrabi et al., 2021).
- **Assessment Tools** An interesting direction that researchers have taken is introducing tools that can assess the amount of fairness in a tool or system. For example, Aequitas [136] is a toolkit that lets users to test models with regards to several bias and fairness metrics for different population subgroups. Aequitas produces reports from the obtained data that helps data scientists, machine learning researchers, and policymakers to make conscious decisions and avoid harm and damage toward certain populations. AI Fairness 360 (AIF360) is another toolkit developed by IBM in order to help moving fairness research algorithms into an industrial setting and to create a benchmark for fairness algorithms to get evaluated and an environment for fairness researchers to share their ideas [11]. These types of toolkits can be helpful for

learners, researchers, and people working in the industry to move towards developing fair machine learning application away from discriminatory behavior (Mehrabi et al., 2021).

- At first sight, automating decisions may give a sense of fairness: classification rules do not guide themselves by personal preferences. However, at a closer look, one realizes that classification rules are actually learned by the system (e.g., loan granting) from the training data. If the training data are inherently biased for or against a particular community (e.g., foreigners), the learned model may show a discriminatory prejudiced behavior. In other words, the system may infer that just being foreign is a legitimate reason for loan denial. (**M62_**)

- One might think of a straightforward preprocessing approach consisting of just removing the discriminatory attributes from the data set. Although this would solve the direct discrimination problem, it would cause much information loss and in general it would not solve indirect discrimination. (**M62_**)

- Hence, there are two important challenges regarding discrimination prevention: one challenge is to consider both direct and indirect discrimination instead of only direct discrimination; the other challenge is to find a good tradeoff between discrimination removal and the quality of the resulting training data sets and data mining models. (**M62_**)

- Most AI systems and algorithms are data driven and require data upon which to be trained. Thus, data is tightly coupled to the functionality of these algorithms and systems. In the cases where the underlying training data contains biases, the algorithms trained on them will learn these biases and reflect them into their predictions. As a result, existing biases in data can affect the algorithms using the data, producing biased outcomes. Algorithms can even amplify and perpetuate existing biases in the data.(Mehrabi et al., 2021).

- In addition, algorithms themselves can display biased behavior due to certain design choices, even if the data itself is not biased. The outcomes of these biased algorithms can then be fed into real-world systems and affect users' decisions, which will result in more biased data for training future algorithms.(Mehrabi et al., 2021).

- The loop capturing this feedback between biases in data, algorithms, and user interaction is illustrated in Figure 1. We use this loop to categorize definitions of bias in the section below (Mehrabi et al., 2021).

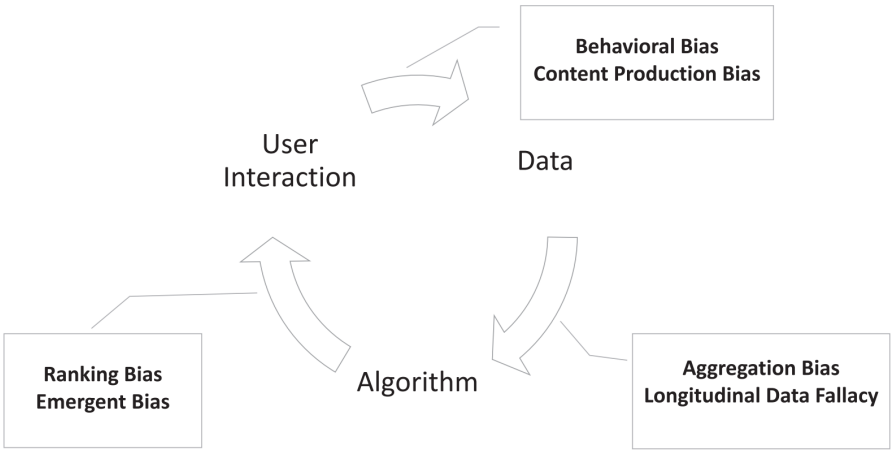

Figure 2.1: Bias definitions in a ML lifecycle (Mehrabi et al., 2021).

- Bias can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks. In (Suresh & Guttag, 2021), authors talk about sources of bias in machine learning with their categorizations and descriptions in order to motivate future solutions to each of the sources of bias introduced in the paper. In (Olteanu et al., 2019), the authors prepare a complete list of different types of biases with their corresponding definitions that exist in different cycles from data origins to its collection and its processing.(Mehrabi et al., 2021).

Already rewritten: The following categorization was modeled with the intent to show that the different biases are intertwined and one should consider the effects between each other in the cycle to address them correctly (Mehrabi et al., 2021)

2.2.3 Bias Overview

Bias	Mentioned in Context of		
	ML	**Dermatology**	**Demography in Dermatology**
Data Biases			
Measurement Bias	X[1,2]		
Omitted Variable Bias	X[1,11,13]		
Representation Bias	X[1,2]		
Aggregation Bias	X[1,2]		
Sampling Bias	X[1]		
Longitudinal Data Fallacy	X[1]		
Linking Bias	X[1,3]		
Algorithmic Biases			
Algorithmic Bias	X[1,4,5]		
User Interaction Bias	X[1,4]		
Presentation Bias	X[1,4]		
Ranking Bias	X[1,4,6]		
Popularity Bias	X[1,10]		
Emergent Bias	X[1,9]		
Evaluation Bias	X[1,2,12]		
User Biases			
Historical Bias	X[1,2]		
Population Bias	X[1,3,8]		
Self-Selection Bias	X[1]		
Social Bias	X[1,4,7]		
Behavioral Bias	X[1,3]		
Temporal Bias	X[1,3]		
Content Production Bias	X[1,3]		
Healthy Volunteer Selection Bias	X[14]		

[1] (Mehrabi et al., 2021)
[2] (Suresh & Guttag, 2021)
[3] (Olteanu et al., 2019)
[4] (Baeza-Yates, 2018)
[5] (Danks & London, 2017)
[6] (Lerman & Hogg, 2014)
[7] (Wang & Wang, 2014)
[8] (Hargittai, 2007)
[9] (Friedman & Nissenbaum, 1996)
[10] (Ciampaglia et al., 2018)
[11] (Clarke, 2005; Riegg, 2008)
[12] (Buolamwini & Gebru, 2018)
[13] (Mustard, 2003)
[14] (**M54_**)

Table 2.3: Biases - Mentioned in Contextual Research, grouped like in Mehrabi et al., 2021, the author cannot guarantee for completeness

Bias sensitive Features	Mentioned in Context of		
	ML	**Dermatology**	**Demography in Dermatology**
Affected Features			
Skin Type	X[1,2,8]		
Gender/Sex	X[1,2,8,9,10,11,12]		
Gender and Skin Type Subgroups	X[1,2]		
Geographic Diversity	X[1,3]		
Ethnicity/Race TODO: check definitions	X[1,2,4,5,6,7,8,12]		
Socio-Economic Status	X[7]		
Familial status	X[8]		
Disabilities	X[8,12]		
Marital status	X[8,12]		
Recipient of public assistance	X[8]		
Age	X[8,12]		
Religion	X[8,12]		
Nationality/National origin	X[8,12]		

[1] (Mehrabi et al., 2021) [5] (**M143_**) [9] (**M167_**)
[2] (Buolamwini & Gebru, 2018) [6] (**M54_**) [10] (**M20_**)
[3] (**M142_**) [7] (**M150_**) [11] (**M168_**)
[4] (**M98_**) [8] (**M30_**) [12] (**M62_**)

Table 2.4: Features which often hold biases - Mentioned in Contextual Research, grouped like in Mehrabi et al., 2021, the author cannot guarantee for completeness

2.2.4 Biases Extensive Sources

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

- Measurement Bias. Measurement, or reporting, bias arises from how we choose, utilize, and measure particular features (Suresh & Guttag, 2021) (e.g. mismeasured proxy variables (= "one or more variables that encode the protected attribute with a substantial degree of accuracy" according to https://medium.com/bcggamma/practice-ai-responsibly-with-proxy-variable-detection-42c2156ad986)) (Mehrabi et al., 2021). (= e.g. someone who lives at that postal code probably has this ethnicity https://medium.com/bcggamma/practice-ai-responsibly-with-proxy-variable-detection-42c2156ad986); –> could that be an issue with the country of origin feature?

- Omitted Variable Bias. Omitted variable bias4 occurs when one or more important variables are left out of the model (Clarke, 2005; Riegg, 2008)(Mustard, 2003). Something that the model was not ready for(Mehrabi et al., 2021). did not take into account

- Representation Bias. Representation bias arises from how we sample from a population during data collection process (Suresh & Guttag, 2021). Non-representative samples lack the diversity of the population, with missing subgroups and other anomalies (Mehrabi et al., 2021).

- Aggregation Bias. Aggregation bias (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population. An example of this type of bias can be seen in clinical aid tools. Consider diabetes patients who have apparent morbidity differences across ethnicities and genders. Specifically, HbA1c levels, that are widely used to diagnose and monitor diabetes, differ in complex ways across genders and ethnicities. Therefore, a model that ignores individual differences will likely not be well-suited for all ethnic and gender groups in the population (Suresh & Guttag, 2021). This is true even when they are represented equally in the training data. Any general assumptions about subgroups within the population can result in aggregation bias. (Mehrabi et al., 2021). –> could also be important for dermatology issues!!!

  - Simpson's Paradox. Simpson's paradox is a type of aggregation bias that arises in the analysis of heterogeneous data [18]. The paradox arises when an association observed in aggregated data disappears or reverses when the same data is disaggregated into its underlying subgroups (Fig. 2(a)). ... After analyzing graduate school admissions data, it seemed like there was bias toward women, a smaller fraction of whom were being admitted to graduate programs compared to their male counterparts. However, when admissions data was separated and analyzed over the departments, women applicants had equality and in some cases even a small advantage over men. The paradox happened as women tended to apply to departments with lower admission rates for both genders. Simpson's paradox has been observed in a variety of domains, including biology [37], psychology [81], astronomy [109], and computational social science [91].(Mehrabi et al., 2021).
  - Modifiable Areal Unit Problem is a statistical bias in geospatial analysis, which arises when modeling data at different levels of spatial aggregation [56]. This bias results in different trends learned when data is aggregated at different spatial scales (Mehrabi et al., 2021).

- Sampling Bias. Sampling bias is similar to representation bias, and it arises due to nonrandom sampling of subgroups. As a consequence of sampling bias, the trends estimated for one population may not generalize to data collected from a new population. (Mehrabi et al., 2021). This is what the PASSION dataset tries to improve

- Longitudinal Data Fallacy. Researchers analyzing temporal data must use longitudinal analysis to track cohorts over time to learn their behavior. Instead, temporal data is often modeled using cross-sectional analysis, which combines diverse cohorts at a single time point. The heterogeneous cohorts

can bias cross-sectional analysis, leading to different conclusions than longitudinal analysis (Mehrabi et al., 2021). –> could this be relevant for the progress of a specific disease? Or would that only be an issue when the progress of the disease would be predicted?

- Linking Bias. Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users (Olteanu et al., 2019) (Mehrabi et al., 2021). –> probably less important since we got individuals? Or could that be an issue with the country of origin feature?

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

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

- User Interaction Bias. User Interaction bias is a type of bias that can not only be observant on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction (Baeza-Yates, 2018). This type of bias can be influenced by other types and subtypes, such as presentation and ranking biases. (Mehrabi et al., 2021). – more relevant for later, when the application would become bigger

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

- Popularity Bias. Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation—for example, by fake reviews or social bots (Ciampaglia et al., 2018). ... this presentation may not be a result of good quality; instead, it may be due to other biased factors. (Mehrabi et al., 2021).

- Emergent Bias. Emergent bias occurs as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values, or societal knowledge usually some time after the completion of design (Friedman & Nissenbaum, 1996). This type of bias is more likely to be observed in user interfaces, ... This type of bias can itself be divided into more subtypes, as discussed in detail in (Friedman & Nissenbaum, 1996). (Mehrabi et al., 2021). probably less relevant at the first stage

- Evaluation Bias. Evaluation bias happens during model evaluation (Suresh & Guttag, 2021). This includes the use of inappropriate and disproportionate benchmarks for evaluation of applications such as Adience and IJB-A benchmarks. These benchmarks are used in the evaluation of facial recognition systems that were biased toward skin color and gender (Buolamwini & Gebru, 2018), and can serve as examples for this type of bias (Suresh & Guttag, 2021). (Mehrabi et al., 2021). – important for this thesis

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

- Historical Bias. Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection (Suresh & Guttag, 2021). ... search results were of course reflecting the reality, but whether or not the search algorithms should reflect this reality is an issue worth considering (Mehrabi et al., 2021) - maybe relevant

- Population Bias. Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population (Olteanu et al., 2019). Population bias creates non-representative data. ... More such examples and statistics related to social media use among young adults according to gender, race, ethnicity, and parental educational background can be found in (Hargittai, 2007). (Mehrabi et al., 2021)

- Self-Selection Bias. Self-selection bias4 is a subtype of the selection or sampling bias in which subjects of the research select themselves. (Mehrabi et al., 2021)

- Social Bias. Social bias happens when others' actions affect our judgment (Baeza-Yates, 2018). (case where we want to rate or review an item with a low score, but when influenced by other high ratings, we change our scoring thinking that perhaps we are being too harsh [(Baeza-Yates, 2018), (Wang & Wang, 2014).) (Mehrabi et al., 2021)

- Behavioral Bias. Behavioral bias arises from different user behavior across platforms, contexts, or different datasets (Olteanu et al., 2019). (Mehrabi

et al., 2021) maybe, people from different countries go to the dermatologist for different diseases, based on cultural differences?

- Temporal Bias. Temporal bias arises from differences in populations and behaviors over time (Olteanu et al., 2019). (Mehrabi et al., 2021) – could this also be differences in the year, when people go to dermatologists? over which timeline has the PASSION data been captured?
- Content Production Bias. Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contents generated by users (Olteanu et al., 2019). (Mehrabi et al., 2021) – could the quality of the pictures been related to this as well?

2.2.5 Fairness Overview

Fairness Definitions	**Mentioned in Context of**		
	ML	**Dermatology**	**Demography in Dermatology**
Group Fairness			
Conditional Statistical Parity	X[1,3,10]		
Demographic/Statistical Parity	X[1,3,4,5]		
Equalized Odds	X[1,2,3]		
Equal Opportunity	X[1,2,3]		
Treatment Equality	X[1,7]		
Test Fairness	X[1,3,8]		
Subgroup Fairness			
Subgroup Fairness	X[1,11,12]		
Individual Fairness			
Fairness Through Awareness	X[1,4,5]		
Fairness Through Unawareness	X[1,5,6]		
Counterfactual Fairness	X[1,5]		
Not Categorized			
Fairness in Relational Domains	X[1,9]		

[1] (Mehrabi et al., 2021)
[2] (Hardt et al., 2016)
[3] (Verma & Rubin, 2018)
[4] (Dwork et al., 2012)
[5] (Kusner et al., 2017)
[6] (Grgic-Hlača et al., 2016)
[7] (Berk et al., 2017)
[8] (Chouldechova, 2017)
[9] (Farnadi et al., 2018)
[10] (Corbett-Davies et al., 2017)
[11] (Kearns et al., 2018)
[12] (Kearns et al., 2019) TODO: potential bias in this bc same author of the algorithm tested it

Table 2.5: Fairness Definitions - Mentioned in Contextual Research, grouped like in Mehrabi et al., 2021, the author cannot guarantee for completeness

2.2.6 Fairness Extensive Sources

Algorithmic Fairness

- in order to be able to fight against discrimination and achieve fairness, one should first define fairness. (Mehrabi et al., 2021)
- The fact that no universal definition of fairness exists shows the difficulty of solving this problem [138]. Different preferences and outlooks in different cultures lend a preference to different ways of looking at fairness, which makes it harder to come up with just a single definition that is acceptable to everyone in a situation. there is still no clear agreement on which constraints are the most appropriate for those problems. (Mehrabi et al., 2021)
- Broadly, fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [139]. Even though fairness is an incredibly desirable quality in society, it can be surprisingly difficult to achieve in practice. (Mehrabi et al., 2021)
- Here we will reiterate and provide some of the most widely used definitions, along with their explanations inspired from (Verma & Rubin, 2018).(Mehrabi et al., 2021)
- Definitions on page 12, 13, 14 (Mehrabi et al., 2021)
  - Equalized Odds (TP and FP rate should be the same for individuals in different sub groups) (Mehrabi et al., 2021)
  - Equal Opportunity (TP rate should be the same) (Mehrabi et al., 2021)
  - Demographic Parity / Statistical Parity (likelihood of positive outcome the same regardless of protected group) (Mehrabi et al., 2021)
  - Fairness Through Awareness (similar predictions to similar individuals (similarity = inverse distance)) (Mehrabi et al., 2021)
  - Fairness Through Unawareness (no protected attributes explicitly used in decision-making process) (Mehrabi et al., 2021)
  - Treatment Equality (Ration FN and FP same for both protected group categories) (Mehrabi et al., 2021)
  - Test Fairness (for predicted probability scores, people in both groups must have equal probability of TP) (Mehrabi et al., 2021)
  - Counterfactual Fairness (same outcome in actual world and counterfactual world where the individual belonged to a different demographic group) (Mehrabi et al., 2021)
  - Fairness in Relational Domains (“A notion of fairness that is able to capture the relational structure in a domain—not only by taking attributes of individuals into consideration but by taking into account the social, organizational, and other connections between individuals”

(Farnadi et al., 2018)) (Mehrabi et al., 2021) probably not relevant since not relational

- Conditional Statistical Parity (people in both groups have equal possibilities of being assigned to a positive outcome given a set of legitimate factors) (Mehrabi et al., 2021)
- My text: The survey categorizes those fairness notions in three different groups: Individual Fairness, Group Fairness and Subgroup fairness. (Mehrabi et al., 2021)
- Subgroup fairness: Subgroup fairness intends to obtain the best properties of the group and individual notions of fairness. It is different than these notions but uses them in order to obtain better outcomes. It picks a group fairness constraint like equalizing false positive and asks whether this constraint holds over a large collection of subgroups (Kearns et al., 2018)(Kearns et al., 2019)(Mehrabi et al., 2021)
- it is impossible to satisfy some of the fairness constraints at once except in highly constrained special cases. In [83], the authors show the inherent incompatibility of two conditions: calibration and balancing the positive and negative classes. These cannot be satisfied simultaneously with each other unless under certain constraints; therefore, it is important to take the context and application in which fairness definitions need to be used into consideration and use them accordingly [141](Mehrabi et al., 2021)
- Another important aspect to consider is time and temporal analysis of the impacts that these definitions may have on individuals or groups. In [95] authors show that current fairness definitions are not always helpful and do not promote improvement for sensitive groups—and can actually be harmful when analyzed over time in some cases. They also show that measurement errors can also act in favor of these fairness definitions; therefore, they show how temporal modeling and measurement are important in evaluation of fairness criteria and introduce a new range of trade-offs and challenges toward this direction. It is also important to pay attention to the sources of bias and their types when trying to solve fairness-related questions. (Mehrabi et al., 2021)

2.2.7 Mitigation Methods Overview

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

Mitigation Methods - Unbiasing Data (Pre-Processing)	**Mentioned in Context of**		
	ML	**Dermatology**	**Demography in Dermatology**
Good Practices while using Data	X[1,2,3]		
Datasheets as supporting document for dataset creation method, characteristics, motivations and skews	X[1,2,3]		
Datasheets as supporting document for model method, characteristics, motivations and skews	X[1,4]		
Dataset Nutrition Label	X[1,5,6]		
Test for Simpson's Paradox TODO: Discribe Simpson's Paradox	X[1,7,8,9]		
Detect Direct Discrimination with Causal Models and Graphs	X[1,10]		
Preventing Direct and Indirect Discrimination	X[1,11]		
Messaging	X[1,12]		
Preferential Sampling	X[1,13,14]		
Disparate Impact Removal	X[1,15]		

[1] (Mehrabi et al., 2021)
[2] (**M13_**)
[3] (**M55_**)
[4] (**M110_**)
[5] (**M66_**)
[6] (**M66Successor_**)
[7] (**M81_**)
[8] (**M3_**)
[9] (**M4_**)
[10] (**M163_**)
[11] (**M62_**)
[12] (**M74_**)
[13] (**M75_**)
[14] (**M76_**)
[15] (**M51_**)

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

Mitigation Methods - Fair Classification	Mentioned in Context of		
	ML	Dermatology	Demography in Dermatology
satisfy certain fairness definitions			
Satisfy subgroup fairness TODO: unclear if * in [3] as well, or if [2] also handles *	X[1,2]		
Satisfy equality of opportunity*	X[1,3,6]		
Satisfy equalized odds*	X[1,3]		
Disparate Treatment**	X[1,4,5]		
Disparate Impact**	X[1,4,5]		
TODO: find out what	X[1,7]		
TODO: find out what	X[1,8]		
TODO: find out what	X[1,9]		
TODO: find out what	X[1,10]		
satisfy fairness definitions and stability for test set changes			
TODO: find out what	X[1,11]		
Adaptions of existing classifiers			
Modified discrimination-free Naive Bayes classifier	X[1,12]		
Frameworks			
Framework for fairness-aware classification	X[1,13]		
Fairness constraints in Multitask learning (MTL) framework	X[1,14]		
Decoupled Classification System with Transfer Learning	X[1,15]		
Preferential Data Usage TODO: find better name			
Wasserstein Distance Measure for mitigating dependence on sensitive attributes	X[1,16]		
Preferential Sampling (PS) for discrimination free train data set	X[1,17]		
Provide Interpretability			
Post-Processing with attention mechanism	X[1,18]		

* 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_**)

Table 2.7: Mitigation Methods - Fair Classification - Mentioned in Contextual Research, grouped like in Mehrabi et al., 2021, the author cannot guarantee for completeness

Mitigation Methods - not so relevant for us	Mentioned in Context of		
	ML	Dermatology	Demography in Dermatology
Fair Representation Learning (Pre/In-processing TODO: check with Ludovic)			
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]		
Fair NLP TODO: check with Ludovic			
Fair Word-Embedding	X[1,5,6,7]		
Train-Time Data Augmentation	X[1,8]		
Test-Time Neutralization	X[1,8]		
Fair Regression (In-processing)			
Price of Fairness (POF)	X[1,10]		
XY TODO: check this and bounded group loss	X[1,11]		
Decision Tree for Disoarate 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]		
Others TODO: find other categorization			
Disparate Learning Processes (DLP)	X[1,9]		
Disregard sensitive attributes in effect on decision making	X[1]		
Community Detection / Graph Embedding TODO: how to proceed with this	X		
Causal Approach to Fairness TODO: how to proceed with this	X		
Disregard path in causal graph which result in sensitive attributes affecting decision outcome	X[1]		

[1] (Mehrabi et al., 2021)
[2] (**M42_**)
[3] (**M97_**)
[4] (**M112_**)
[5] (**M20_**)
[6] (**M58_**)
[7] (**M169_**)
[8] (**M166_**)
[9] (**M94_**)
[10] (**M14_**)
[11] (**M1_**)
[12] (**M2_**)
[13] (**M167_**)
[14] (**M137_**)
[15] (**M5_**)
[16] (**M90_**)
[17] (**M65_**)

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

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

2.2.8 Mitigation Methods Extensive Sources

Bias Examples and Mitigation Ideas Data bias examples and mitigation ideas

- Bias in ML Data - (Buolamwini & Gebru, 2018) IJB-A / Adience imbalanced (mainly light-skinned subjects) - Bias towards dark-skinned groups (under-represented). Other instance - when we do not consider different subgroups in the data. Considering only male-female groups not enough, use race to further subdivide gender groups. Only then, clear biases in sub groups can be found, since otherwise part of the groups would compromise the other group and hide the underlaying bias towards that subgroup (Mehrabi et al., 2021)

- Popular machine-learning datasets that serve as a base for most of the developed algorithms and tools can also be biased—which can be harmful to the downstream applications that are based on these datasets. ... In [142], researchers showed that these datasets suffer from representation bias and advocate for the need to incorporate geographic diversity and inclusion while creating such datasets. (Mehrabi et al., 2021)

- Examples of Data Bias in Medical Applications. These data biases can be more dangerous in other sensitive applications. For example, in medical domains there are many instances in which the data studied and used are skewed toward certain populations—which can have dangerous consequences for the underrepresented communities. [98] showed how exclusion of African-Americans resulted in their misclassification in clinical studies, so they became advocates for sequencing the genomes of diverse populations in the data to prevent harm to underrepresented populations (Mehrabi et al., 2021) TODO: What does sequencing data mean?, is it relevant

- Authors in [143] studied the 23andMe genotype dataset and found that out of 2,399 individuals, who have openly shared their genotypes in public repositories, 2,098 (87%) are European, while only 58 (2%) are Asian and 50 (2%) African (Mehrabi et al., 2021)

- Other such studies were conducted in [54] which states that UK Biobank, a large and widely used genetic dataset, may not represent the sampling population. Researchers found evidence of a “healthy volunteer” selection bias. [150] has other examples of studies on existing biases in the data used in the medical domain. [157] also looks at machine-learning algorithms and data utilized in medical fields, and writes about how artificial intelligence in health care has not impacted all patients equally.(Mehrabi et al., 2021)

Methods for Fair Machine Learning

- While this section is largely domain-specific, it can be useful to take a cross-domain view. Generally, methods that target biases in the algorithms fall under three categories (Mehrabi et al., 2021)
- Pre-processing. Pre-processing techniques try to transform the data so that the underlying discrimination is removed [43]. If the algorithm is allowed to modify the training data, then pre-processing can be used [11].(Mehrabi et al., 2021)
- In-processing. In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process [43]. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used during the training of a model— either by incorporating changes into the objective function or imposing a constraint [11, 14].(Mehrabi et al., 2021)
- Post-processing. Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model [43]. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase [11, 14].(Mehrabi et al., 2021)
- we concentrate on discrimination prevention based on preprocessing, because the preprocessing approach seems the most flexible one: it does not require changing the standard data mining algorithms, unlike the inprocessing approach, and it allows data publishing (rather than just knowledge publishing), unlike the postprocessing approach. (**M62_**) –> TODO: this is an important point which we should consider for PASSION, also, some more insight in regards of the different phases can be found in this paper
- From learning fair representations [42, 97, 112] to learning fair word embeddings [20, 58, 169], debiasing methods have been proposed in different AI applications and domains. (Mehrabi et al., 2021) –> seems to refer mostly to NLP domains
- Most of these methods try to avoid unethical interference of sensitive or protected attributes into the decision-making process, while others target exclusion bias by trying to include users from sensitive groups. (Mehrabi et al., 2021)
- However, a recent paper [58] argues against these debiasing techniques and states that many recent works on debiasing word embeddings have been superficial, that those techniques just hide the bias and don't actually remove it. (Mehrabi et al., 2021)

- some works try to satisfy one or more of the fairness notions in their methods, such as disparate learning processes (DLPs) which try to satisfy notions of treatment disparity and impact disparity by allowing the protected attributes during the training phase but avoiding them during prediction time [94].(Mehrabi et al., 2021)

- Some of the existing work tries to treat sensitive attributes as noise to disregard their effect on decision-making, while some causal methods use causal graphs, and disregard some paths in the causal graph that result in sensitive attributes affecting the outcome of the decision.(Mehrabi et al., 2021)

- Different bias-mitigating methods and techniques are discussed below for different domains—each targeting a different problem in different areas of machine learning in detail. (Mehrabi et al., 2021)

Unbiasing Data

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

- Authors in [66] also propose having labels, just like nutrition labels on food, in order to better categorize each data for each task. (Mehrabi et al., 2021)

- some work has targeted more specific types of biases. For example, [81] has proposed methods to test for cases of Simpson's paradox in the data, and [3, 4] proposed methods to discover Simpson's paradoxes in data automatically. (Mehrabi et al., 2021)

- Causal models and graphs were also used in some work to detect direct discrimination in the data along with its prevention technique that modifies the data such that the predictions would be absent from direct discrimination [163].(Mehrabi et al., 2021)

- in [62] also worked on preventing discrimination in data mining, targeting direct, indirect, and simultaneous effects.(Mehrabi et al., 2021)

- Other pre-processing approaches, such as messaging [74], preferential sampling [75, 76], disparate impact removal [51], also aim to remove biases from the data. (Mehrabi et al., 2021)

Fair Classification

- certain methods have been proposed [57, 78, 85, 106] that satisfy certain definitions of fairness in classification. For instance, in [147] authors try to satisfy subgroup fairness in classification, equality of opportunity and equalized odds in [63], both disparate treatment and disparate impact in [2, 159], and equalized odds in [154]. (Mehrabi et al., 2021)
- Other methods try to not only satisfy some fairness constraints but to also be stable toward change in the test set [69] (Mehrabi et al., 2021)
- The authors in [155], propose a general framework for learning fair classifiers. This framework can be used for formulating fairness-aware classification with fairness guarantees. In another work [25], authors propose three different modifications to the existing Naive Bayes classifier for discrimination-free classification.(Mehrabi et al., 2021)
- paper [122] takes a new approach into fair classification by imposing fairness constraints into a Multitask learning (MTL) framework. In addition to imposing fairness during training, this approach can benefit the minority groups by focusing on maximizing the average accuracy of each group as opposed to maximizing the accuracy as a whole without attention to accuracy across different groups. In a similar work [49], authors propose a decoupled classification system where a separate classifier is learned for each group. They use transfer learning to reduce the issue of having less data for minority groups.(Mehrabi et al., 2021)
- In [73] authors propose to achieve fair classification by mitigating the dependence of the classification outcome on the sensitive attributes by utilizing the Wasserstein distance measure.(Mehrabi et al., 2021)
- In [75] authors propose the Preferential Sampling (PS) method to create a discrimination free train data set. They then learn a classifier on this discrimination free dataset to have a classifier with no discrimination.(Mehrabi et al., 2021)
- In [102], authors propose a post-processing bias mitigation strategy that utilizes attention mechanism for classification and that can provide interpretability. (Mehrabi et al., 2021)

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

- "price of fairness" (POF) to measure accuracy-fairness trade-offs, 3 penalites: Individual fairness, group fairness and hybrid fairness [14] (Mehrabi et al., 2021)
- In addition to the previous work, [1] considers the fair regression problem formulation with regards to two notions of fairness statistical (demographic)

parity and bounded group loss. [2] uses decision trees to satisfy disparate impact and treatment in regression tasks in addition to classification. (Mehrabi et al., 2021)

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

- RBA (reducing bias amplification) as calibration algorithm to prevent risk of leveraging social bias, distributions in training data are followed in the predictions. multi-label obeject and visual semantic role labeling classification amplify existing bias in data [167] (Mehrabi et al., 2021) –> TODO: be careful with this if the approach would be to generate new images for training!!

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

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

Community Detection/Graph Embedding This types of algorithms are mainly used on data in regards of connections between the data, e.g. in online communities and social networks. Please refer to Mehrabi et al., 2021 for more information, as this type of data is not found in the PASSION context.

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

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

- Variational Auto encoders –> Variational Fair Autoencoder introduced in [97]. Here,they treat the sensitive variable as the nuisance variable, so that by removing the information about this variable they will get a fair representation. They use a maximum mean discrepancy regularizer to obtain invariance in the posterior distribution over latent variables. Adding this

maximum mean discrepancy (MMD) penalty into the lower bound of their VAE architecture satisfies their proposed model for having the Variational Fair Autoencoder.
In [5] authors also propose a debiased VAE architecture called DB-VAE which learns sensitive latent variables that can bias the model (e.g., skin tone, gender, etc.) and propose an algorithm on top of this DB-VAE using these latent variables to debias systems like facial detection systems.
In [112] authors model their representation-learning task as an optimization objective that would minimize the loss of the mutual information between the encoding and the sensitive variable. The relaxed version of this assumption is shown in Equation 1. They use this in order to learn fair representation and show that adversarial training is unnecessary and in some cases even counter-productive.
In [42], authors introduce flexibly fair representation learning by disentanglement that disentangles information from multiple sensitive attributes. Their flexible and fair variational autoencoder is not only flexible with respect to downstream task labels but also flexible with respect to sensitive attributes. They address the demographic parity notion of fairness, which can target multiple sensitive attributes or any subset combination of them. (Mehrabi et al., 2021)

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

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

- Word Embedding TODO: potentially relevant, if the labels are used in training, e.g. age / gender determines how the disease is presenting itself in the images; check (Mehrabi et al., 2021) page 21 if relevant

- Coreference Resolution "Coreference resolution involves identifying when two or more expressions in a text refer to the same entity, be it a person, place, or thing." https://medium.com/@datailm/the-key-to-unlocking-true-language-understanding-coreference-resolution-c01d569e2e87 TODO: irrelevant for the PASSION Context

comparison of different mitigation algorithms

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

2.3 Statistical biases

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

- Selection bias - wrong sampling method, working on a specific subset of audience; usually by working only with data that is easy to access
- Self-selection bias - when you let the subjects of the analyses select themselves, less proactive people will be excluded TODO: could be an issue as well for PASSION, couldn't it? since the doctors probably ask the clients. One way to go is to default should be to provide access to the data. but is it ethical?
- Recall bias - respondent doesn't remember things correctly TODO: keep an eye on this when recalling evidences!!
- Observer bias - projecting expectations onto the research
- Survivorship bias
- Omitted variable bias
- Cause-effect bias
- Funding bias
- Cognitive bias

2.4 Dermatology Bias

- https://ijdvl.com/biases-in-dermatology-a-primer/ 29 biases, 4 reasons to know about it, 7 mitigation methods

3 Ideas and Concepts

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

3.1 PASSION Dataset

TODO: write things to consider more precisely:

- Include more details in gender attribute - transgender have probably different genes / hormones, and should be indicated for more accurraccy
- include profession / at least an adapted version to indicate high risk patients for certain diseases? -> might lead to other biases?
- change country of origin to ethnicity (less of a proxy variable)
- are the data collectors specialized in some fields? That could lead to bias towards the center's country and the diagnosed diseases
- include images of healthy skin

3.2 Broad Methodology

TODO: write things to consider more precisely:

- Divide and Conquer vs. All-In-One-Model
  - An algorithm per ethnicity / subgroup running at the same time
  - Running 1 Algorithm chosen based on Fitzpatrick skin type
  - Running 1 Algorithm which detects first the demographic subgroup (FST, gender, age, ...) and runs the specific subgroup algorithm afterwards

  - Hint Ludovic: Still not of data, maybe also others; often limited because the data is missing, you are missing data from others

- BLIND performance vs. Including the demographic data
  - Idea to try if the labels are not relevant for the diagnosis and should only be used for evaluating fairness purposes as some papers suggest
  - Might be obsolete after demographic biases in dermatology research, since melanin response and melanoma risk is different in male and female according to research https://pmc.ncbi.nlm.nih.gov/articles/PMC4797181/

- Hint Ludovic: Maybe Focal Loss more relevant –> emphasis on data vs model

- Divide and Conquer vs. All-In-One-Model (either by etnicity x algorithms at a time or one which seperates the imgs first by demographic subgroup (incl. Fitzpatrick skin type))

- BLIND performance vs. Including the demographic data

4 Methods

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

5 Execution

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

6 Evaluation and Validation

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

7 Outlook

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

8 Glossary

TODO: Add List of Formulas if necessary TODO: add AI declarations somewhere

9 Bibliography

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

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art [Publisher: SAGE Publications Inc]. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533
Mehrabi 15.

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).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments [Publisher: Mary Ann Liebert, Inc., publishers]. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047
Mehrabi 34.

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.

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

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

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

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.

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.

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.

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.

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

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

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

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

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.

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.

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.

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

TODO: check the gls all unused.