Demographic Biases in Dermatology Models

Nadja Stadelmann

March 8, 2025

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: BSc Computer Science

Year of Graduation: 2025

Main Advisor: Ludovic Amruthalingam

External Expert:

Industry partner/provider: Applied AI Research Lab, Lucerne University of Applied Sciences and Arts

Code / Thesis Classification:
☒ Public (Standard)
☐ Private
TODO: fix linebreaks and indents

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 TODO: add thanks and gratitude
Ludovic Amruthalingam
Simone Lionetti - deputy Ludovic
Pascal Baumann - LaTeX

TODO: Your abstract here.

Contents

Todo list

TODO: use the template on overleaf: https://www.overleaf.com/project/67cac1b71993d0b44b6ba7ee
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

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

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

- 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).
- 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).

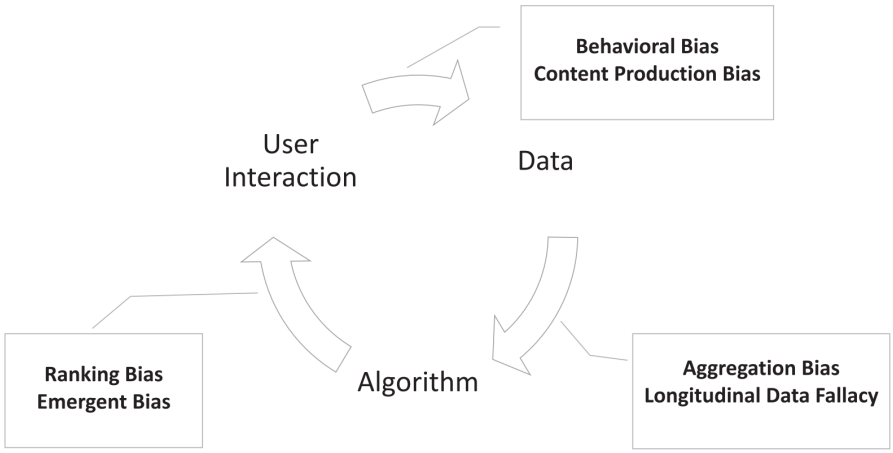

Figure 2.1.: Bias definitions in a ML lifecycle (Mehrabi et al., 2021).

- Bias can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks. In [144], 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 [120], 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).
- (Mehrabi et al., 2021).
- (Mehrabi et al., 2021).
- (Mehrabi et al., 2021).
- (Mehrabi et al., 2021).
- (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)

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 [144] (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/bcggam 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 [38, 114, 131]. 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 [144]. 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 [144]. 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 [120] (Mehrabi et al., 2021). –> probably less important since we got individuals? Or could that be an issue with the country of origin feature?

- (Mehrabi et al., 2021).

- (Mehrabi et al., 2021).

- (Mehrabi et al., 2021).

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 [9]. 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 [44], 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 [9]. 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 [9] (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 [9] and crowdsourcing applications [93].(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 [117]. ... 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 [53]. 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 [53]. (Mehrabi et al., 2021). probably less relevant at the first stage

- Evaluation Bias. Evaluation bias happens during model evaluation [144]. 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 [24], and can serve as examples for this type of bias [144]. (Mehrabi et al., 2021). – important for this thesis

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 [144]. ... 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 [120]. 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 [64]. (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 [9]. (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 [9, 151].) (Mehrabi et al., 2021)
- Behavioral Bias. Behavioral bias arises from different user behavior across platforms, contexts, or different datasets [120]. (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 [120]. (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 [120]. (Mehrabi et al., 2021) – could the quality of the pictures been related to this as well?
- (Mehrabi et al., 2021)
- (Mehrabi et al., 2021)
- (Mehrabi et al., 2021)
- (Mehrabi et al., 2021)
- (Mehrabi et al., 2021)

Data bias examples and mitigation ideas

- Bias in ML Data - IJB-A / Adience imbalanced (mainly light-skinned subjects) - Bias towards dark-skinned groups (underrepresented). Other instance - when we do not consider different subgroups in the data. Considering only male-female groups not enough, use race to further subdivide gender groups. Only then, clear biases in sub groups can be found, since otherwise part of the groups would compromise the other group and hide the underlaying bias towards that subgroup (Mehrabi et al., 2021)
- Popular machine-learning datasets that serve as a base for most of the developed algorithms and tools can also be biased—which can be harmful to the downstream applications that are based on these datasets. ... In [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

- 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)

- (Mehrabi et al., 2021)

- (Mehrabi et al., 2021)

Discrimination vs. bias

- 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)

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 [149].(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" [50]) (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 [79][80](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)

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

TODO: move on with page 13-15

2.2.1. ML biases

2.2.2. ML fairness metrics

2.2.3. ML 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

3.2. Broad Methodology

TODO: write things to consider more precisely:

- 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

FST Fitzpatrick skin type, skin classifier based on the skins' reaction to light (Gottfrois et al., 2024). 5

Jupyter Notebook Executable files, often used in ML to write Python code and add explanations in text form.. 8

pediatric A medical term for infants, children and adolescents.. 5

TODO: Add to ToC of content somehow and fix chapter numbers

List of Figures

List of Tables

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

9. Bibliography

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/

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.

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.

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

A. Appendix

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.