Universiteit Leiden

Analyzing Representation and Bias in Face Recognition Datasets & The Role of Documentation in Enhancing Transparency

Thulashikaa Vinasiththamby (s3858774)

Supervisors:
Zane Kripe & Kristoffer Kalavainen

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 1, 2026

Abstract

Facial recognition systems are increasingly deployed in high-stakes contexts. While these systems are often evaluated in terms of overall accuracy, evidence suggests that their performance can vary across demographic groups. These disparities are closely linked to the datasets used to train such systems. This thesis investigates representational bias in three facial recognition datasets: MORPH-II, VGGFace2, and Balanced Faces in the Wild. In addition, the accompanying documentation is examined to assess how dataset composition, limitations, and potential biases are communicated. The findings reveal substantial demographic imbalances in MORPH-II and VGGFace2, whereas BFW is balanced by design. Furthermore, analysis shows that dataset documentation often fails to critically reflect on the implications of these imbalances, even describing datasets as diverse despite significant underrepresentation of certain groups. This limits transparency and makes it difficult for users to assess whether a dataset is suitable for a particular context of use. The findings highlight the importance of greater awareness of representational bias, more consistent documentation practices, and increased transparency in dataset reporting to support the development of fairer facial recognition systems.

Contents

1 Introduction

Data has become a integral part of everyday life, shaping the systems and structures around us. As data is collected, transformed into datasets, then used to train models, and finally deployed in real world systems, biases present at early stages can persist throughout the entire process. These biases may influence decisions that have real consequences for the lives of individuals.

In particular, facial recognition datasets represent a critical area of concern. Facial images are often collected in specific social contexts and later reused to train models, often with limited transparency about how and why they were originally gathered. Well-known incidents highlight the risks of such systems. For example, Joy Buolamwini's experience offers a striking illustration of this problem when a facial recognition system failed to detect her face until she put on a white mask [7]. Similarly, Google Photos mislabeled photos of Black individuals as "gorillas" in 2015 [4]. These cases are shown in Figure 1. These incidents reveal how biased or insufficient training data can encode and reproduce inequalities at scale. When such systems are integrated into institutional settings, these errors are not merely technical limitations; they can lead to misidentification, discrimination, and lasting social harm to people's reputations and opportunities. This shows a lack of algorithmic fairness, meaning systems do not perform equally across different demographic groups.

Figure 1: (Left) Joy Buolamwini with a white mask because a facial recognition system failed to detect her face. (Right) Google Photos mislabeled photos of Black individuals as "gorillas".

However, data is often treated as an abstract or purely technical object rather than a socially situated object, shaped by the historical, cultural, and political contexts in which they are produced. Decisions about who is included, who is excluded and what is measured are not neutral choices; they reflect existing power structures and social inequalities. Abstraction, the process of reducing complex social categories categories into fixed labels, while presenting them as neutral and objective, can conceal systematic patterns of underrepresentation or misrepresentation. This allows structural inequalities to persist unnoticed within the dataset itself. Attention to bias and representation is crucial for producing fairer models [14].

While datasets have been accompanied by documentation, this information has often focused on technical characteristics rather than issues such as data collection practices, representation, and potential sources of bias. Growing concerns about bias, fairness, and accountability in machine learning systems have led to calls for more systematic and transparent documentation practices.

Such improved documentation practices can make dataset composition and limitations more explicit [19]. For instance, if documentation clearly indicates that certain demographic groups are underrepresented, this information can inform critical decisions about whether a dataset is suitable for high-stakes applications such as facial recognition in security or identity verification contexts.

In addition to supporting dataset users, documentation practices can also influence dataset creators themselves. The introduction of structured documentation frameworks can encourage creators to reflect on their own assumptions and potential biases during the dataset construction process [19].

However, it remains unclear how consistently such documentation practices are implemented in practice and how effectively they surface representational biases. Therefore, a closer examination of both dataset composition and its documentation is needed to assess how well these biases are actually made visible and addressed. This thesis investigates representational biases in face recognition datasets and examines the ways in which these biases are documented and communicated.

The main research question guiding this study is: *What representational biases exist in facial recognition datasets, and how does dataset documentation reflect these biases?* In order to address this question, the study first examines what types of representational biases can be identified in facial recognition datasets, including imbalances across demographic attributes such as race, gender, and age. It then investigates what kind of documentation is provided alongside these datasets, and to what extent this documentation reveals or omits information about such biases.

1.1 Thesis overview

This bachelor thesis was conducted in the Data Science & Artificial Intelligence program at LIACS, Leiden University, under the supervision of Zane Kripe and Kristoffer Kalavainen.

The thesis is structured as follows: Section 2 provides background on bias, representation, facial recognition datasets and documentation practices. Section 3 describes the research approach and methodology. Section 4, 5 & 6 presents the dataset analysis and results. Section 7 discusses the findings. Section 8 provides reflections on data science education. Finally, Section 9 concludes the thesis and outlines directions for further research.

2 Background

2.1 Data Is Not Neutral

The machine learning field has placed large-scale datasets at the center of model development. The dominant framing within computer science continues to treat data-driven methods as objective and impartial. However, this assumption has been challenged across a range of fields, including critical data studies, science and technology studies and feminist and sociological approaches to technology. These perspectives argue that data are not neutral representations of reality, but are instead socially constructed and often reflect existing patterns of discrimination, historical prejudices, and unequal social structures [6][14][35][42].

Machine learning systems are particularly susceptible to reproducing such biases because they learn patterns directly from historical data. Even when developers do not intentionally encode discriminatory rules, algorithms can inherit and amplify biases present in the training data [1].

Consequently, algorithmic discrimination can emerge, challenging the notion that automated systems are inherently neutral or objective. Large-scale evaluations conducted by the U.S. National Institute of Standards and Technology (NIST), as discussed in the Harvard Journal of Law & Technology review on racial bias in facial recognition, show that many commercial algorithms are between 10 and 100 times more likely to misidentify Black or East Asian faces than white faces. In some cases, American Indian faces are the most frequently misidentified, while Black women experience the highest error rates overall. These disparities are not random, but are systematically linked to the conditions under which systems are developed and trained [8][24]. Other research shows how datasets in fields such as health research, safety testing, and product design frequently treat male bodies as the default standard. As a result, systems and protocols built on such data often perform less effectively for women [13].

Comparable issues have also been identified in remote assessment technologies. Research on remote proctoring systems shows that these tools are less accurate for certain groups of students, particularly in relation to skin tone and disability. Face detection and tracking components tend to perform worse on students with darker skin tones, including more frequent errors in detecting facial features and a higher rate of false positives, where normal behavior is incorrectly flagged as suspicious. Similar problems occur for students with disabilities. For example, students with autism, whose eye contact patterns may differ from those assumed by the system or students with motor impairments that may move in ways that differ from expected behavioral models, can all be disproportionately flagged because the systems are trained on normative behavioral patterns [9]. In both cases, these errors stem from design assumptions embedded in training data and system development, where system performance and expectations are implicitly calibrated toward narrow and exclusionary normative standards.

These issues are commonly summarized by the principle of "garbage in, garbage out"[3]. Machine learning models rely on training data as ground truth, meaning that the quality and representativeness of the data directly shape model behavior. When datasets contain biased, incomplete, or inaccurate information, the resulting systems may produce unreliable predictions and discriminatory outcomes. Rather than eliminating bias, algorithmic systems can therefore perpetuate and legitimize existing inequalities by presenting their outputs as data-driven and objective [3][14]. This highlights the importance of "decentering technology" in discussions of discrimination [17]. This means shifting attention away from algorithms as isolated objects and toward the wider systems of power and social relations in which they are embedded. Only by doing so can responses to discrimination address underlying causes rather than just their technical manifestations [17][42].

Beyond societal bias reflected in data, datasets are also shaped by a series of design and collection choices. Decisions about what counts as data, how categories are defined, which variables are recorded, and which populations are made visible or rendered invisible are not neutral technical steps. Instead, they reflect underlying values of dataset creators, institutional priorities, and broader relations of power. These choices can ultimately affect the fairness and performance of

machine learning systems [14].

2.1.1 Power Structures in Machine Learning Development

The broader power structures in which machine learning algorithms are developed and deployed are often overlooked [42]. They are typically controlled by organizations with significant economic and political influence, giving them substantial authority over how these systems are designed, implemented and used [12][23]. Questions of representation, fairness, and accountability are therefore closely connected to questions of who has the authority to collect data, define categories, and determine how machine learning systems are deployed [12][42].

The concentration of decision-making power also affects how the benefits and harms of systems are distributed. Organizations that develop and deploy the systems may gain efficiency and profit, while the consequences of biased or inaccurate systems are often experienced by individuals with limited influence over their design [14]. This raises important questions about whose interests are prioritized during the development process and who has the ability to challenge or influence these decisions [12].

D'Ignazio and Klein situate these issues within broader systems of oppression, which involve the systematic mistreatment of certain groups by others [14]. Oppression emerges when power is not distributed equally and one group gains disproportionate control over institutions such as law and education, and uses its power to systematically exclude others while granting unfair advantages to themselves, often reinforcing existing social inequalities [3].

An important challenge in recognizing such inequalities is the concept of *privilege hazard* [14]. Individuals who benefit from existing power structures are often less able to recognize the ways those structures disadvantage others. This lack of lived experience can limit their ability to identify instances of oppression, anticipate potential harms, or imagine alternative solutions. As a result, privilege hazards are not necessarily the product of intentional discrimination but can arise from a lack of awareness about how systems affect different groups [46].

These dynamics are particularly relevant in the creation of datasets. Dataset creators are not always aware of the assumptions and biases embedded in their work, a phenomenon described as the *power paradox*, in which those with the greatest influence over data systems may be least aware of their impacts [14]. This lack of awareness affects dataset construction [44].

2.2 Representation and Biases

Bias in computer systems can be understood as systematic and unfair discrimination against certain individuals or groups in favor of others. A system is biased when it denies opportunities or assigns undesirable outcomes to individuals or groups on grounds that are unreasonable or inappropriate [15].

Bias can arise at multiple stages of the machine learning pipeline. During data collection, sampling bias occurs when the collected data do not adequately represent the target population, resulting in some groups being overrepresented while others are underrepresented. During dataset construction, selection bias may emerge through decisions about which data are included or

excluded. Annotation bias can arise when labels assigned by human annotators are shaped by subjective interpretations or social stereotypes. Although these biases originate at different stages of the data pipeline, they ultimately influence how individuals are represented within datasets [29][54].

Representation is a fundamental issue in the development of machine learning systems. Datasets do not merely reflect reality but actively shape how reality is represented. The importance of representation extends beyond technical performance and can reproduce and amplify existing social inequalities [44]. As Crawford argues, the greatest threat posed by artificial intelligence is not that machines will become more intelligent than humans, but that they will encode sexism, racism, and other forms of discrimination into the digital infrastructure of society [12][14].

Representation within datasets plays a crucial role in determining the fairness and performance of machine learning systems. When certain demographic groups are included less frequently than others, underrepresentation occurs, which can result in poorer performance for those groups because the model has fewer examples from which to learn.

Conversely, overrepresentation can also introduce distortions in the learned data distribution. Stereotype-aligned correlations between demographic attributes and depicted activities have been identified in several computer vision datasets, for example where women are overrepresented in images depicting cooking and shopping [52]. Such patterns indicate that overrepresentation may reflect not only frequency differences, but also socially structured patterns of visibility. Overrepresentation can further lead to systematically skewed observations. Barocas et al. note that when certain groups are subject to disproportionate attention, their actions are more likely to be observed and recorded, which can result in higher recorded rates of mistakes or negative outcomes compared to less scrutinized groups [3].

It is also important to consider how different dimensions of identity intersect with one another [52]. The concept of intersectionality highlights that characteristics such as race, gender and age do not operate independently. Instead, these dimensions interact to shape an individual's experiences and exposure to discrimination. As a result, bias cannot always be understood by examining race or gender separately. For example, a facial recognition system may perform adequately on average for women and for people with darker skin tones, while still performing poorly for darker-skinned women specifically. In such cases, the bias arises from the combination of race and gender rather than from either characteristic alone. Ignoring these intersections can therefore hide important disparities and lead to an incomplete understanding of bias in machine learning systems [7][14].

2.3 Facial Recognition Datasets

Computer vision is a subfield of machine learning concerned with enabling systems to interpret and extract meaningful information from visual data, such as images and videos, in order to make predictions. One of the most widely studied applications within computer vision is face recognition. The development and evaluation of face recognition systems depend heavily on large-scale datasets containing visual data, typically ranging from tens of thousands to millions of facial images. These datasets typically consist of collections of images accompanied by annotations that provide information about each data instance [52].

Several characteristics make facial recognition datasets unique within the machine learning field. Faces encode a large number of discriminative features that are highly individual and difficult to anonymize. Unlike many other data types, faces are biometric and directly tied to personal identity. Because faces are personally identifiable, their collection and use raise ethical concerns related to privacy, consent and potential misuse. Even when data is sourced from public images, individuals may not expect their faces to be used in algorithms [57].

Research has shown that facial recognition systems do not always perform equally for all populations. For example, research by Klare et al. revealed that the accuracy of face recognition systems used by US-based law enforcement is systematically lower for people labeled Female or Black [36]. Facial recognition technologies are increasingly used in high-stakes contexts, including law enforcement. More than 117 million Americans are included in law enforcement face recognition networks. A year-long research investigation across 100 police departments revealed that African Americans are more likely to be subjected to facial recognition searches by law enforcement than people with other ethnicities, increasing their exposure to potential misidentification, false positives and unwarranted searches [8][18].

In the widely cited study *Gender Shades*, Buolamwini and Gebru found substantial disparities in performance across gender [8]. They demonstrated that male subjects were generally classified more accurately than female subjects. They further found that lighter-skinned individuals were generally classified more accurately than darker-skinned individuals. An intersectional breakdown revealed that the poorest performance was consistently observed for darker-skinned females in particular. While lighter-skinned males experienced very low error rates, darker-skinned females were misclassified at significantly higher rates, with error rates of up to 34.7% for darker-skinned females, compared to a maximum error rate of only 0.8% for lighter-skinned males [8]. These results demonstrate that performance disparities are not random but follow consistent patterns across demographic groups [14].

2.4 Documentation practices

The manner in which dataset developers choose to present and describe their work has consequences for how their dataset is perceived and impacts the dataset use. Studies show that the field of machine learning has not traditionally developed a strong culture of reflexivity regarding values and assumptions embedded in their work [33][46]. As a result, the presentation of datasets as objective artifacts contributes to a culture of uncritical and unquestioning dataset use among machine learning practitioners [52].

Datasets that power machine learning are often used, shared, and reused with little visibility into the processes of deliberation that led to their creation [27]. As machine learning systems are increasingly deployed in high-stakes tasks, greater transparency about data is needed and stronger accountability for the decisions made and the development of the dataset [31].

As Heinke et al. note, prevailing cultural norms among dataset generators prioritize speed and rapid progress toward improved benchmark performance, sometimes at the expense of careful consideration of data reuse, data management, and legal or ethical implications [27]. Importantly,

the challenges associated with datasets are not solely due to their invisibility to end users, but also stem from the systematic devaluation of dataset creation as a reflective and socially situated process, since dataset creation is treated as a purely technical step [27][31][52].

Improving the responsible development of machine learning systems, through reflexivity on assumptions and design choices and transparency regarding potential implications for algorithmic fairness, requires that datasets are documented in a clear and structured manner [52]. This involves providing information about a dataset's intended purpose, data collection and composition, the motivation behind its creation and the contexts in which it should or should not be used. Documentation practices serve two primary stakeholder groups: dataset creators and dataset consumers. For dataset consumers, documentation helps to better understand the dataset's limitations and to identify possible sources of bias [2][27]. In this way documentation can influence model building decisions, because when model developers are aware of demographic imbalances or limitations in datasets, they can make more informed choices about whether it is suitable for a particular context of use. Transparency on the part of dataset creators is necessary for dataset consumers to avoid unintentional misuse. For dataset creators, documentation can encourage them to reflect on their own assumptions and potential biases. Importantly, documentation can create a loop for dataset creators in which awareness of bias can lead to changes in data collection and dataset design, such as actively collecting additional data for underrepresented groups or rebalancing the dataset, ultimately contributing to improved accountability and more responsible dataset development [19].

To this end, several dataset documentation frameworks have been proposed to promote transparency and accountability, including Datasheets for Datasets [19], Dataset Nutrition Labels [11][28], and Data Cards [45]. These approaches share a common goal: to encourage dataset generators to reflect carefully on the process of creating a dataset, and to equip dataset consumers with the information needed to make informed use of a dataset for their machine learning tasks [27].

However, the three approaches differ in both emphasis and structure [27]. The *Datasheets for Datasets* approach was inspired by datasheets used in the electronics industry, which are used to describe the properties of hardware components. Gebru et al. structure documentation as answers to a set of guiding questions that prompt dataset creators to systematically record important aspects of dataset design, collection, and intended use [19].

The *Dataset Nutrition Label* takes inspiration from the idea of food nutrition labels, aiming to present key dataset information in a concise and easily interpretable format, similar to how nutritional content is summarized for food products [11][28].

The *Data Card* framework is designed to support clear and standardized documentation of datasets for a range of stakeholders, including ML practitioners, product teams, and other decision-makers. It provides a flexible and modular structure that can be adapted to different datasets and application contexts, enabling more consistent communication about dataset composition, collection, and intended use [45].

These differences suggest that existing dataset documentation frameworks are largely complementary, each being better suited to particular contexts, audiences, or stages of the data lifecycle. At the same time, the variation in structure, scope, and level of detail across approaches

highlights the absence of a single, universally adopted standard for dataset documentation in machine learning practice [27][31].

Findings indicate that although dataset documentation frameworks are widely cited and there is broad recognition of their importance for supporting responsible ML development, their actual adoption in practice remains limited and in many cases close to absent [27].

This limited real-world adoption can likely be explained by a combination of incentive, infrastructural, regulatory and cultural factors. There are currently few, if any, direct career or funding incentives associated with producing detailed, transparent dataset documentation, making its adoption largely voluntary. In addition, creating comprehensive documentation requires significant time, expertise, and organizational resources, which can act as a substantial barrier to implementation. Regulatory guidance in this area also tends to be non-binding, offering recommendations rather than enforceable requirements, which further reduces pressure for adoption. Finally, dominant cultural norms within machine learning practice often prioritize rapid experimentation, model performance, and publication speed over careful documentation [27][58].

The absence of comparable requirements for dataset documentation in major computer vision conferences may contribute to the limited visibility of dataset creation practices and quality assessments. Current submission guidelines for major computer vision conferences, including CVPR, ICCV, and ECCV do not currently actively enforce or require structured dataset documentation as part of supplementary materials. This lack of emphasis suggests a limited focus on systematically measuring and documenting broader aspects of dataset quality, such as representational balance and associated limitations, within the computer vision research community [58].

3 Approach

The research combines quantitative and qualitative analysis of three datasets. Quantitative analysis makes it possible to identify and measure demographic imbalances, while qualitative analysis enables an assessment of whether these imbalances are acknowledged in the accompanying documentation.

3.1 Method for dataset analysis

First, for each dataset, the dataset composition is examined by looking at the gender, race and age group frequency count. Besides that, intersectional combinations (e.g., gender within ethnic categories) were analyzed because previous research has shown that in some cases, disparities in facial recognition systems emerge at the intersection of demographic characteristics (e.g., Asian Female) rather than within a single characteristic alone (e.g., Black) [8].

The three face recognition datasets are selected based on the accessibility of both the data and accompanying documentation for analysis. This is important because it ensures that both the dataset and documentation can be properly analyzed. Besides that, they were also chosen based on the diversity of design contexts. The three selected datasets are MORPH-II, VGGFace2, and Balanced Faces in the Wild (BFW). In particular, BFW was mentioned to be specifically designed to achieve demographic balance, in contrast to MORPH-II and VGGFace2, which were

not constructed with this goal in mind. This difference in design context allows for an interesting comparison between datasets collected for different purposes and a dataset explicitly designed with demographic balance in mind.

During the dataset selection process, other face recognition datasets were also considered but were excluded because the data was either not publicly accessible, had no available labels that could be found, or lacked an accompanying documentation paper.

MORPH II is one of the largest publicly available facial image datasets, made at the University of North Carolina. It compromises of more than 55,000 images from more than 13,000 subjects. The MORPH-II dataset is widely utilized in research on gender and race classification, as well as age estimation and synthesis [5]. It is comprised of mug shots taken of arrested people between 2003 and late 2007 [59].

The VGGFace2 dataset was developed by researchers at the Visual Geometry Group (VGG) at the University of Oxford and released in 2018. It contains approximately 3.31 million images of 9,131 identities, primarily consisting of celebrity photographs collected through Google Image Search. The dataset was designed to capture a wide range of variations in pose, age and lighting, making it suitable for training deep learning models for face recognition. The dataset has widely been used for face recognition, age recognition and identity verification [10].

Balanced Faces in the Wild (BFW) is a specialized face recognition dataset, constructed to be balanced across demographic subgroups. The BFW dataset was introduced by Robinson et al. (2023), with several of the authors affiliated with Northeastern University [51]. BFW contains 800 faces from 100 unique subjects. Each subject contributes 25 images. Subjects are grouped into eight demographic subgroups based on ethnicity (e.g., Asian, Black, Indian, White) and gender (female/male). Because BFW was specifically designed with demographic balance in mind, it serves as a useful benchmark for evaluating how dataset design choices can impact facial recognition systems.

3.1.1 Demographic distribution

The dataset is imported into a pandas DataFrame from a CSV file. An initial inspection of the first rows is performed to verify that the columns (race, gender, and age) were correctly loaded. If the dataset encodes race and gender in abbreviated form, these are mapped to meaningful labels using dictionary mappings: e.g., "W" is replaced with "White" and "B" with "Black" for race, while "M" is replaced with "Male" and "F" with "Female" for gender. To compute the distribution of race and gender, the function `value_counts(normalize=True)` is applied to each column. This operation counts the number of occurrences of each category and divides by the total number of rows, producing relative frequencies. These values are then multiplied by 100 to express them as percentages.

Since the dataset records age as a continuous variable, ages were grouped into four intervals: 18–29, 30–39, 40–49, and 50+. These categories were chosen to make the analysis easier to interpret and to allow comparisons between broader age groups rather than individual ages. This is achieved using `pd.cut()`, which assigns each individual to an age category based on their value. The distribution of these age groups is then computed as percentages of the full dataset using `value_counts(normalize=True)`, allowing comparison across age groups.

To better understand potential demographic imbalances within the dataset, an intersectional analysis was conducted between *gender (Male/Female)* and *race (White/Black)*. The analysis was implemented using a cross-tabulation (`pd.crosstab`) between the variables *race* and *gender*. The resulting contingency table was normalized by the total number of samples and converted into percentages to improve interpretability. The resulting matrix, therefore, represents the proportion of a specific race–gender combination (e.g., Black Male, White Female).

Mathematically, the computation can be expressed as:

$$Percentage_{i,j} = \frac{n_{i,j}}{N} \times 100$$

where $n_{i,j}$ denotes the number of samples in race group i and gender group j, and N represents the total number of samples in the dataset.

3.2 Method for documentation analysis

Dataset documentation refers to the publication papers introducing the dataset [47][10][51]. This study employs documentation analysis to examine the documentation accompanying each face recognition dataset. These are reviewed for every dataset to assess transparency, completeness, and the extent to which creators acknowledge limitations or biases. Comparative analysis across the three datasets is then used to identify differences in how datasets are constructed and how bias and representation are handled in their documentation.

4 MORPH-II

4.1 Dataset Analysis

The MORPH-II dataset with images was found on Kaggle, an online platform widely used for sharing datasets. On Kaggle, the dataset has only a short description, which highlights that it is a pre-processed version with images already cropped and aligned using DLIB and OpenCV [43]. The focus is on convenience and readiness for model training. There is little contextual information about the dataset's origin, collection process, or potential limitations, making it difficult to assess its reliability. Labels (only age and gender) are provided in CSV files.

Figure 2: Sample images from the MORPH-II dataset across different subjects.

However, I also wanted to analyze the race labels, and these were not included in the Kaggle version. Because of this, I searched further and found race labels on the github repository

maintained by Yiming Lin [40]. The github repository is linked to an academic paper about face parsing (FP) for age [41]. In the FP-Age paper, Lin et al. mention MORPH-II as a benchmark dataset for facial age estimation, used to evaluate a deep learning method called face parsing attention [41].

Both the dataset on Kaggle and the labels from GitHub were freely accessible. There were no explicit paywalls or warning materials. Notably, none of the sources provide any ethical considerations or warnings, despite the sensitivity of biometric and demographic data.

Some concerns can be identified with the Kaggle dataset. The Kaggle page does not provide documentation beyond the raw files, limiting transparency about the dataset's background. There is no clear explanation of where or how the dataset was sourced. Additionally, only age and gender labels are included, suggesting limited demographic data. This restricts the ability to examine representation.

The labels from GitHub also raise important issues. First, race is labeled in a highly simplified way. There is only a distinction between Black and White labels. This reduces racial diversity into a binary categorization that does not reflect the broad demographic diversity in the real world. Also, neither the GitHub documentation nor the associated FP-Age paper [38] explains how race labels were originally obtained or verified, again showing a lack of transparency.

4.1.1 Results of Dataset Distribution Analysis

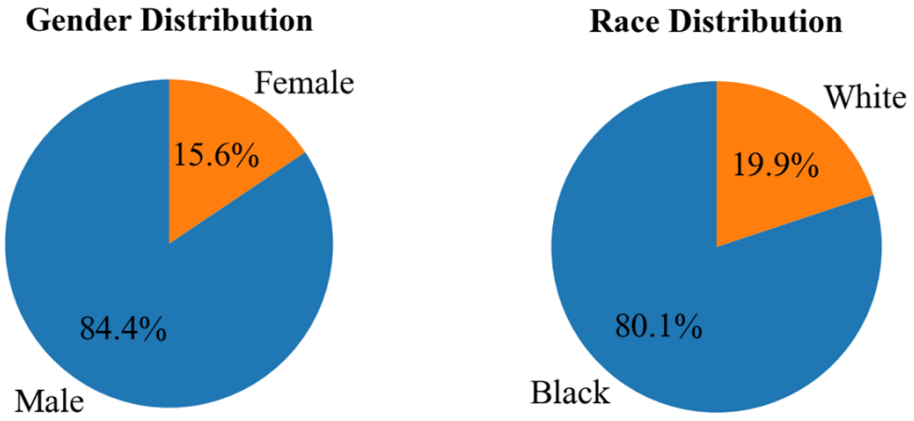

Figure 3: Demographic distribution in the MORPH-II Dataset: Gender (Male/Female) and Race (Black/White) shown as percentages of the total dataset.

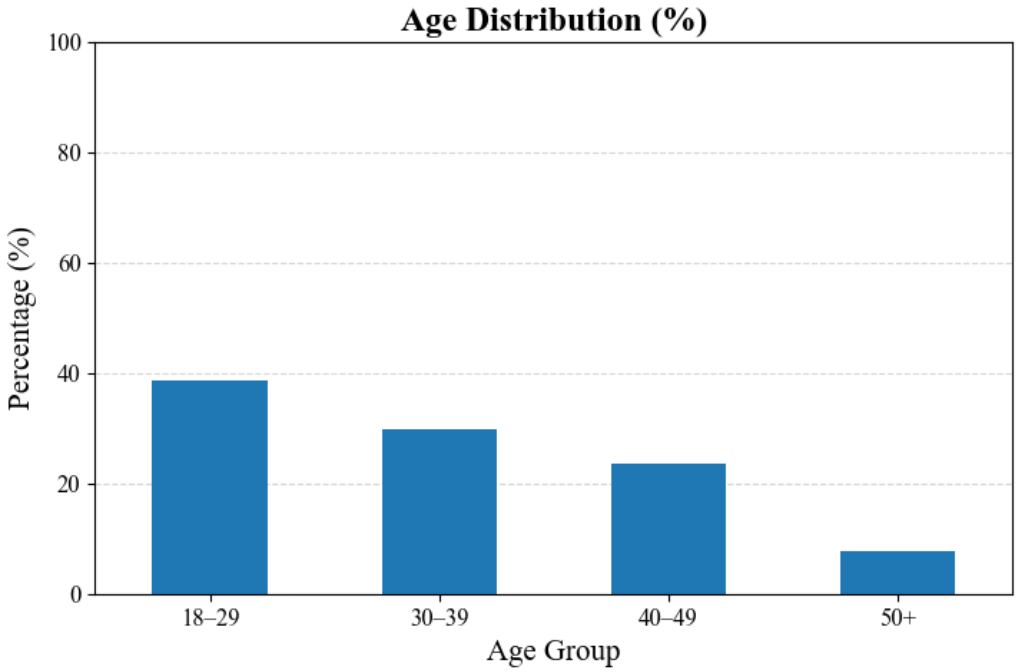

Figure 4: Age distribution of the MORPH-II dataset, grouped into four age categories (18–29, 30–39, 40–49, 50+), shown as percentages of the total dataset.

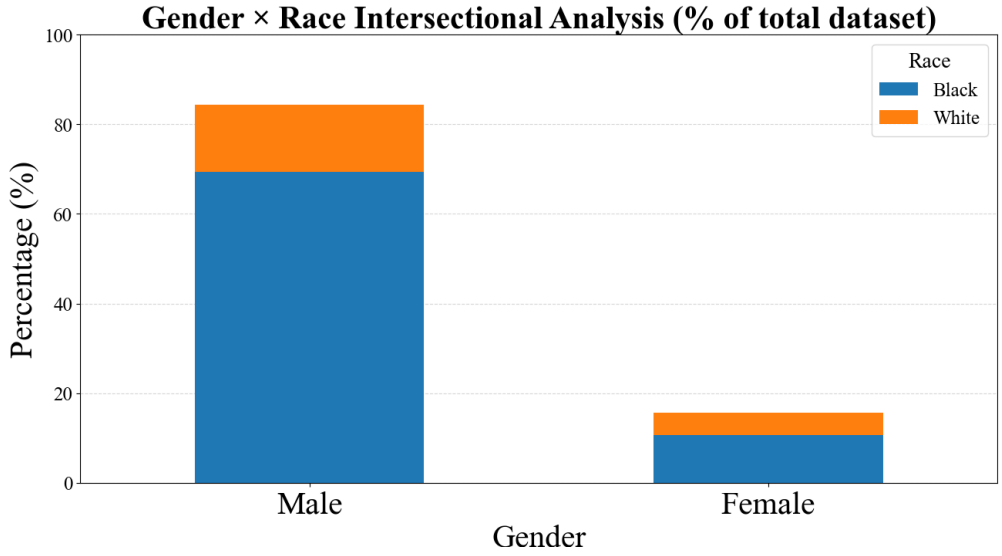

Figure 5: Intersectional analysis of Gender (Male/Female) x Race (Black/White) distributions in the MORPH-II dataset shown as percentages of the total dataset.

4.2 Documentation Analysis

The Kaggle page itself does not provide documentation about the dataset. However, the names of the authors are given on Kaggle. The original documentation for the MORPH dataset was created by K. Ricanek and T. Tesafaye [47]. It is important to note that this documentation refers to the original MORPH dataset, not MORPH-II. MORPH-II is an extended version of the dataset and contains a larger number of images and subjects compared to the original MORPH dataset. There are papers about MORPH-II that are written by researchers associated with the MORPH dataset project and collaborators working within the Face Aging Group at the University of North Carolina Wilmington (UNCW) [37][5].

I was able to find documentation papers of both the MORPH and MORPH-II datasets, though some additional searching was necessary since there were not any papers directly linked on Kaggle. This lack of direct access to dataset information further reduces accessibility and transparency for users attempting to understand the dataset. When searching for the paper, this involved searching Google Scholar for MORPH-II-related documentation and verifying the relevance of the sources by checking how multiple associated studies consistently referenced it as primary documentations when mentioning the MORPH-II dataset.

First, we will begin by examining the documentation of the original MORPH dataset to understand its structure and foundational design, since MORPH-II is built upon it as an extended version. We will then move on to examine the MORPH-II documentation.

4.2.1 MORPH Documentation

The *MORPH: A Longitudinal Image Database of Normal Adult Age-Progression* paper presents MORPH as a longitudinal face image database developed for research in adult age-progression, face modeling, photo-realistic animation, face recognition, etc. [47, p. 1]. It contains duplicate images of the same individuals captured at different ages, allowing researchers to analyze how facial appearance changes over time [47, p. 1].

The database contains demographic attributes such as ethnicity, gender, and age, as well as other data including date of birth (DOB) and date of acquisition (DOA). According to the authors, the images and associated physical properties in MORPH were sourced from public records, with the legacy photographs originally taken between October 26, 1962 and April 7, 1998 [47, p. 2]. Digital scans of these photographs were collected and the authors note that this process was conducted under legal considerations and with approval of the Institutional Review Board (IRB) [47, p. 2]. The reference to IRB approval indicates that the procedures for collecting and handling the images were reviewed under formal institutional ethical guidelines, ensuring that the study met required standards for legal compliance. However, it still raises concerns about how voluntary consent was obtained or interpreted for secondary use of these images in later research and machine learning contexts.

The digital images were cropped about the face to a size of 400 x 500 pixels. The images were converted to grayscale for database uniformity reasons. Some minimal image enhancements were performed on the images; these include artifact minimization with median filtering and contrast stretching via histogram equalization [47, p. 2].

The MORPH database contains 1,724 face images of 515 individuals. The authors claim that

these images represent a diverse population with respect to age, gender, and ethnicity [47, p. 2]. However, they also report that the original MORPH database contains 1,430 images of males and 294 images of females. In addition, they mention it contains 1,278 images of individuals of African-American descent, 433 images of individuals of Caucasian descent and 3 images classified as other [47, p. 2]. This raises questions about the authors' claim of a "diverse population," as the reported numbers indicate that the dataset distribution is highly uneven. While the demographic breakdowns are documented in the paper, the authors do not critically reflect on the representational issues arising from this imbalance, nor do they discuss how such uneven representation may influence model performance or introduce bias.

4.2.2 MORPH-II Documentation

First, we will examine the *MORPH-II: Feature Vector Documentation* by Troy P. Kling from the University of North Carolina Wilmington [37]. Next, we will analyze the *MORPH-II: Inconsistencies and Cleaning Whitepaper*, produced by G. Bingham, B. Yip, M. Ferguson, and C. Nansalo, with guidance from C. Chen, Y. Wang, and T. Kling as part of as part of a National Science Foundation (NSF) Research Experiences for Undergraduates (REU) program at the University of North Carolina Wilmington (Summer 2017) [5].

The *MORPH-II: Feature Vector Documentation* paper is a technical report that describes the structure and features of the MORPH-II facial image dataset [37]. MORPH-II is presented as a database, containing 55,134 mugshots, used for a wide range of purposes, including age estimate, gender and race classification, and facial recognition [37, p. 1]. However, the author mentions there are some challenges with the dataset [37, p. 1]. First, not all images in the original data set have the same resolution. Rather, they come in two varieties: 200x240 pixels and 400x480 pixels (measured width-by-height). The next challenge is that the subjects' distance from the camera varies wildly, with some pictures being taken up close (thereby causing the subject's face to take up the entire frame) and other pictures being taken from far away (thereby causing the subject's face to be much smaller relative to the picture as a whole and surrounded by a blank, uninformative background). This means that a subject's face could occupy a region of the image much smaller than 200x240 pixels. In addition to the subjects' faces being of varying size relative to the image as a whole, many are also tilted to the side by up to 20 or 30 degrees in either direction. The authors note that this results in a majority of images containing faces that are not level, which may impede the performance of machine learning algorithms [37, p. 1]. They further observe that some subjects have their heads tilted toward or away from the camera, introducing another layer of complexity [37, p. 1]. Finally, the lighting conditions in the images are not uniform. Some images are significantly darker than others, and some are so bright that important details of the face are lost. The author says that, ideally, one would want the distribution of light levels to be roughly equal across all images to avoid that changes in illumination affect the accuracy of classification and regression methods [37, p. 2].

This documentation paper primarily frames dataset challenges as technical problems to be addressed through preprocessing and feature engineering, without engaging with issues of representation or potential sources of bias within the dataset. There is little context about the collection

and composition of the dataset itself [37].

The *MORPH-II: Inconsistencies and Cleaning Whitepaper* paper examines the quality and structure of the dataset. Bingham et al. introduce the MORPH-II dataset as a large-scale facial image dataset released in 2008 [5]. It is a collection of 55,134 mugshot images taken between 2003 and late 2007, and includes many images of individuals who were arrested multiple times over these five years. Bingham et al. say this gives the data a longitudinal aspect that has made it very useful in the field of computer vision and pattern recognition and mention in particular that the MORPH-II dataset is widely utilized in research on gender and race classification, as well as age estimation and synthesis [5, p. 1]. According to the local police department, most of the data gathered for mugshots is self-reported, and technology has been used help verify the information being given [5, p. 1].

The paper includes metadata such as date of birth (DOB), date of arrest (DOA), race (categorized as Black, White, Asian, Hispanic, or Other), and gender (Male or Female) [5, p. 2]. Bingham et al. provide a demographic breakdown, mentioning that the dataset includes 36,832 images of male individuals and 5,757 images of female individuals. In terms of race, 42,589 images are classified as Black, 10,559 as White, 154 as Asian, 1,769 as Hispanic, and 63 as Other [5, p. 1].

However, these distributions are presented descriptively rather than critically. Bingham et al. report the numbers, but do not reflect on the implications of such an uneven distribution, nor do they consider how it may influence model behavior. Especially since they themselves mentioned that the MORPH-II dataset is widely utilized in gender and race classification, it is remarkable that they do not address the potential risks associated with training models on a dataset where certain groups (e.g., Black males) are heavily overrepresented, while others (e.g., Asian females) are significantly underrepresented.

Where this paper does go further is in its discussion of data inconsistencies, such as individuals having conflicting labels for race, gender, or date of birth across different entries.

While gender inconsistencies were rare (only one subject was found to have conflicting gender labels), race inconsistencies occurred more frequently, affecting 33 subjects across 132 images. In order to address this and best determine race in the MORPH-II dataset, Bingham et al. applied a cleaning process based on human perception [5, p. 4]. First, literature on race classification was carefully selected and reviewed [16][25][26]. The literature outlines the significance of eyes and nose and the insignificance of features such as skin tone. Researchers were also made aware of possible bias from the other-race-effect: the tendency to more easily recognize faces of one's own race [5, p. 4].

Next, researchers were trained using correctly labelled images from the MORPH-II dataset, with particular attention given to Asian and Hispanic individuals, as these groups accounted for the majority of misclassifications. After reviewing literature and being trained on race perception, the researchers attempted to identify the race of the subjects in question. When a clear majority of images for a given individual belonged to one race, that race was assigned to the subject. In cases where no clear majority existed, annotators' perceived race classification was applied by using the information gathered from the literature and the training acquired from the rest of the dataset. Bingham et al. explicitly note that efforts were made to incorporate multiple perspectives in order to reduce individual bias during the annotation process. This information is particularly

important, as it shows an awareness of the subjectivity involved in human-based race classification and an attempt to mitigate its effects through a more structured and collaborative approach. In cases where the race of an individual could not be determined with sufficient confidence, the subject was assigned to the "Other" category, rather than forcing an uncertain classification [5, p. 4].

Birthdate inconsistencies present a more significant challenge. A total of 1,779 subjects were found to have conflicting birthdate information. While most cases (1,524) could be resolved using a majority rule, the remaining cases involved more complex discrepancies, such as ties or differences of several years. In extreme cases, reported birthdates differed by up to 32 years, and some records resulted in logically inconsistent timelines, such as decreasing age over time. Bingham et al. acknowledge that these issues can significantly impact the performance of models and say that research outcomes could be more accurate if the data were cleaned properly. However, it is also said that there will not likely be an enormous impact on model performance for gender or race prediction, because the number of gender and race inconsistencies is relatively small [5, p. 4].

So interestingly, in this paper bias is briefly acknowledged, specifically in the form of the "other-race effect." However, this is framed as a problem of individual perception that can be mitigated through improved training and standardization of annotators. In this view, bias is treated as something that resides at the level of human judgement and can be corrected through methodological adjustments. There is no consideration of bias as a structural issue embedded in the dataset itself, such as imbalanced demographic representation or the broader context in which the data was collected. As a result, potential biases embedded in the composition of the dataset remain unexamined [5].

5 VGGFace2

5.1 Dataset Analysis

The VGGFace2 dataset was also found on Kaggle. On Kaggle, the dataset is presented as a dataset for face recognition tasks [34]. The description highlights that it contains 3.31 million images of 9,131 identities, with an average of approximately 362.6 images per subject. It is stated that the images were originally collected from Google Image Search, resulting in significant variations in pose, age, illumination, ethnicity, and profession, including public figures such as actors, athletes, and politicians. It is stated that the images are cropped and aligned. Access to the images is provided freely on the Kaggle platform. A significant limitation of the dataset is the complete absence of demographic labels. It does not include any annotations for attributes such as age, gender, or race. Instead, only the raw facial images are provided.

Figure 6: Sample images from the VGGFace2 dataset across different subjects.

In order to conduct a representation analysis of the dataset, it was necessary to search for additional sources with label information about the VGGFace2 dataset. Labels were found on the official VGGFace2 github repository [56]. However, these were limited to gender and a small set of visual attributes such as hair color, facial hair, and the presence of glasses. Importantly, no annotations for age or race were provided within this repository. Consequently, it was necessary to search further for additional sources to obtain age and race labels. Ethnicity labels were obtained from the VGG-Face2 Mivia Ethnicity Recognition (VMER), developed by MIVIA Lab, a research laboratory at the University of Salerno specializing in pattern recognition and computer vision [22]. A link to an accompanying research paper is provided, offering further details on the annotation process [20]. TheVGGFace2 dataset is annotated with 4 ethnicity groups, namely African American (AA), East Asian (EA), Caucasian Latin (CL) and Asian Indian (AI). Greco et al. mention that the concept of ethnicity is rather controversial, since it has been demonstrated that the "ethnicity", as intended by humans, has no biological validity, since there are no genetic characteristics that allow individuals to be grouped according to the commonly defined "ethnicities". However, the categorization is in this case done according to visual differences in the somatic facial features "universally recognized" by humans [20, p. 2].

For instance, the AA group consists of individuals typically having African, North American or South American origins and are characterized by dark skin color, full lips, and wide nose. The EA group belongs to people who have Chinese or other East and South East Asian origins. These individuals are characterized by light skin color, with shades from white to yellowish, and small nose, but the most distinctive feature is the almond shape of the eyes and the inclination between the medial and the lateral canthus, which make the eye look narrower. The CL group consists of individuals that have European, South American, Western Asian and North African origins and are characterized by a white or tanned skin, medium nose and lips and horizontally aligned eyes. The AI group consists of people with Indian, South Asian or Pacific Island origins. They have characteristics in common with EAs and CLs, but are distinguished by noting very slight differences. They have a slightly darker skin color and eyes with more defined features with respect to EAs and CLs [20, p. 5].

Greco et al. mention the other race effect, and explain how it has been scientifically established

that in face analysis tasks humans perform significantly better when dealing with faces of people of their own ethnicity than with individuals belonging to other ethnicities, this effect is also called own race bias. In order to avoid the other race effect, the annotators of the dataset asked three people belonging to different ethnicities, namely one African American, one Caucasian Latin and one Asian Indian, to annotate each identity with an ethnicity label among the considered four [20, p. 4]. To obtain the final annotations, a majority voting rule was applied, which allowed to determine the ethnicity label for 99% of the face images in the dataset; as for the remaining 1%, a tie-break rule was employed, by asking a fourth annotator the opinion about the ethnicity. This fourth annotator had access to additional contextual information, such as the name and background of the individual, which could assist in making a more informed decision [20, p. 6].

The age labels used were also obtained from MIVIA. Greco et al. explain that these labels were generated using an automated prediction approach based on multiple pre-trained models. Instead of relying on a single model, several models were used to estimate the age of each face image. Their predictions were then combined into a single final value, which was assigned as the age label for each image [21, p. 5].

5.1.1 Results of Dataset Distribution Analysis

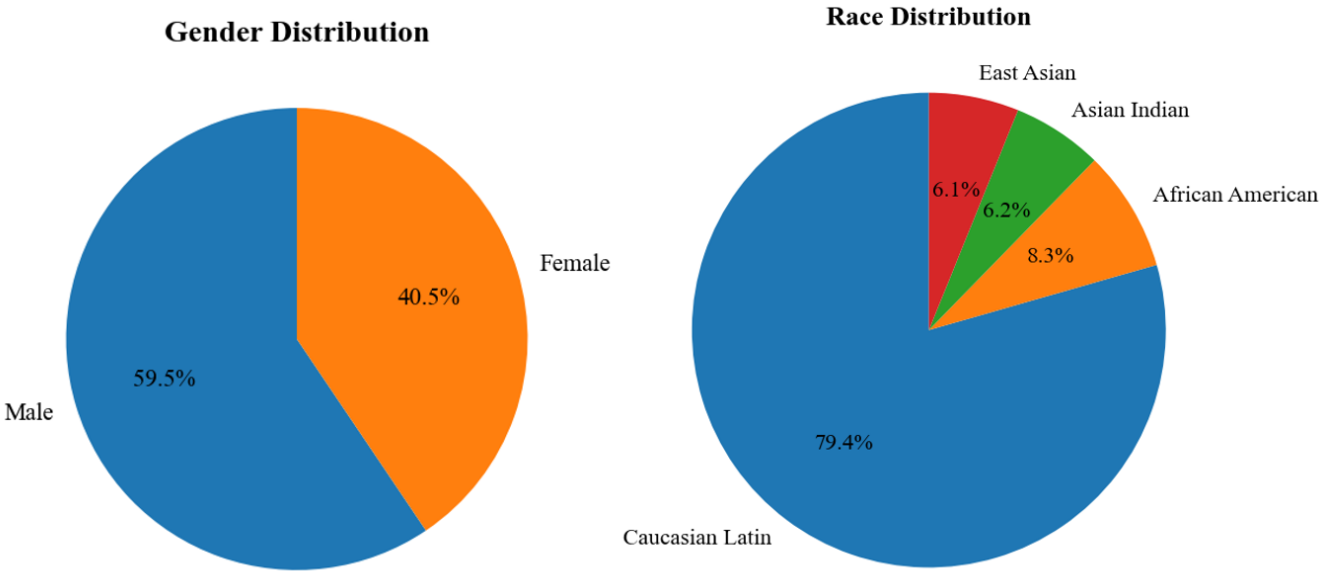

Figure 7: Demographic distribution in the VGGFace2 Dataset: Gender (Male/Female) and Ethnicity (Caucasian Latin, African American, East Asian, Asian Indian) shown as percentages of the total dataset.

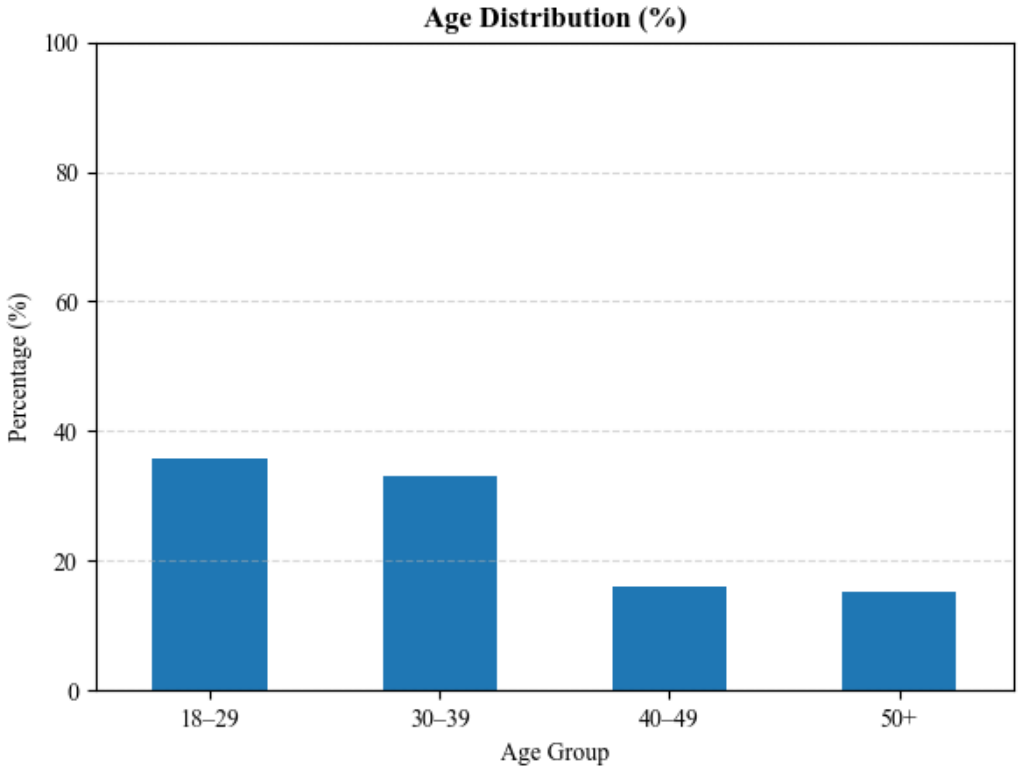

Figure 8: Age distribution of the VGGFace2 dataset, grouped into four age categories (18–29, 30–39, 40–49, 50+), shown as percentages of the total dataset.

5.2 Documentation Analysis

The Kaggle page with the images of the dataset does not provide a link to the associated research paper nor does it reference to the original authors of the dataset. The GitHub repository however includes references to the original documentation of the VGGFace2 dataset authored by Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman from the Visual Geometry Group at the University of Oxford [10].

The original dataset documentation, *VGGFace2: A dataset for recognizing faces across pose and age* describes large-scale face dataset containing 3.31 million images of 9,131 subjects. The images are downloaded from Google Image Search and are described as having large variations in pose, age, illumination, ethnicity and profession (e.g., actors, athletes, politicians). First, as many subjects as possible are found that have a sufficiently distinct set of images available, for example, celebrities and public figures. An initial list of 500k public figures was obtained from the Freebase knowledge graph. An annotator team was then used to remove identities from the candidate list that did not have enough distinct images. For each of the 500K names, 100 images were downloaded using Google Image Search and human annotators were instructed to only keep individuals for which approximately 90% or more of the 100 images belonged to a single identity. This filtering process removed candidates who did not have sufficient images, as well as cases where a single name referred to multiple people in the Google Image Search results. In this manner, the

candidates list was reduced to only 9,244 names. Attribute information such as ethnicity and kinship is obtained from DBPedia. Queries were performed in Google Image Search and 1000 images were downloaded for each subject. To obtain images with large pose and age variations, additional queries were performed by adding terms such as "sideview" and "very young" to each subject's name, resulting in the download of a further 200 images per query. This resulted in 1400 images for each identity [10, p. 2].

Cao et al. say the dataset was collected with three goals in mind: to have both a large number of identities and also a large number of images for each identity; to cover a large range of pose, age and ethnicity; and to minimize the label noise [10, p. 1].

The paper mentions multiple times that the dataset contains a wide range of ethnicities [10, p. 1, 2]. However, it is mentioned that the ethnic balance is still limited by the distribution of celebrities and public figures, and professions. But, this is only mentioned once and only in brackets. The paper states that the dataset is approximately gender-balanced [10, p. 1].

6 Balanced Faces in the Wild

6.1 Dataset Analysis

The Balanced Faces in the Wild (BFW) dataset was obtained from the github repository maintained by the original dataset creators [49]. The page includes the original demographic annotations, particularly race and gender labels. Unlike the other two datasets used in this study, namely MORPH-II and VGGFace2, the Balanced Faces in the Wild (BFW) dataset is accompanied by official labels provided by the dataset authors. Well-structured metadata is available with clearly defined labels.

The official annotation file is provided as a dataframe. The key columns of the dataframe are described as follows:

ID represents the row index of the dataframe. *att* refers to the full intersectional demographic attribute labels of the subjects, such as asian_females, black_males, or white_females. *a* represents the abbreviated subgroup label of the subject, such as AF, AM, BF, BM, IF, IM, WF, and WM, where the first letter denotes ethnicity and the second letter denotes gender. *g* represents the gender labels of the subjects: F = Female, M = Male. *e* represents the ethnicity labels: A = Asian, B = Black, I = Indian, W = White.

Table 1: Example metadata entry from the BFW dataset

Field	**Value**	**Explanation**
ID	0	Index (i.e., row number) of dataframe
p	asian_females/n000009/0010_01.jpg	Path of the face image
att	asian_females	Full demographic attribute labels of the subjects (race & gender)
a	AF	Abbreviated subgroup label of gender and race.
g	F	Gender label (F/M)
e	A	Ethnicity label (A, B, I, W)

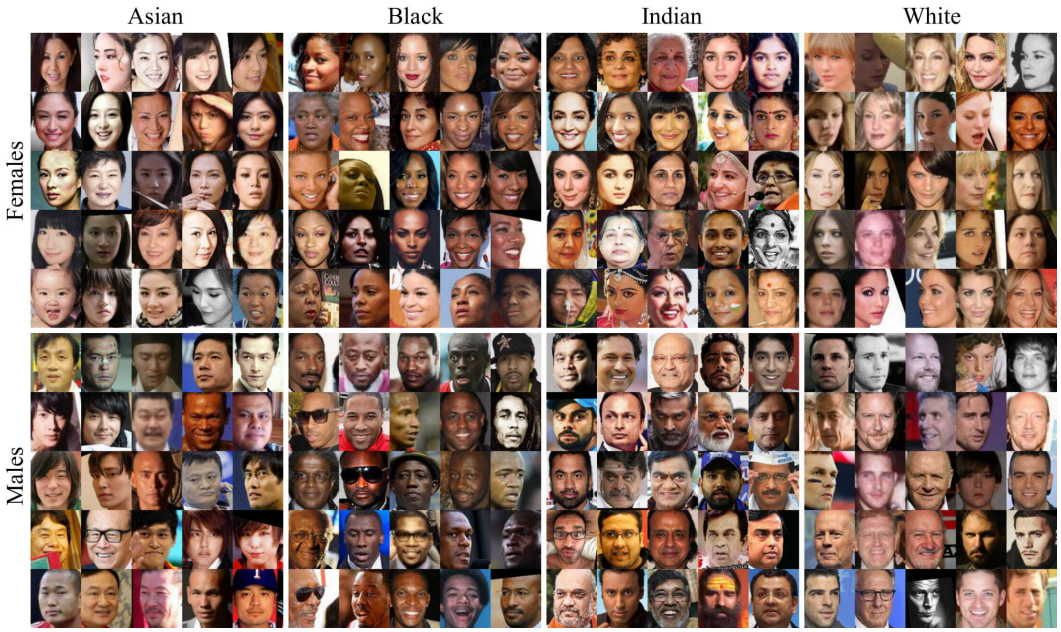

Figure 9: Sample images from the BFW dataset across different subjects from different demographic subgroups (Asian Females, Black Females, Indian Females, White Females, Asian Males, Black Males, Indian Males, White Males).

6.1.1 Results of Dataset Distribution Analysis

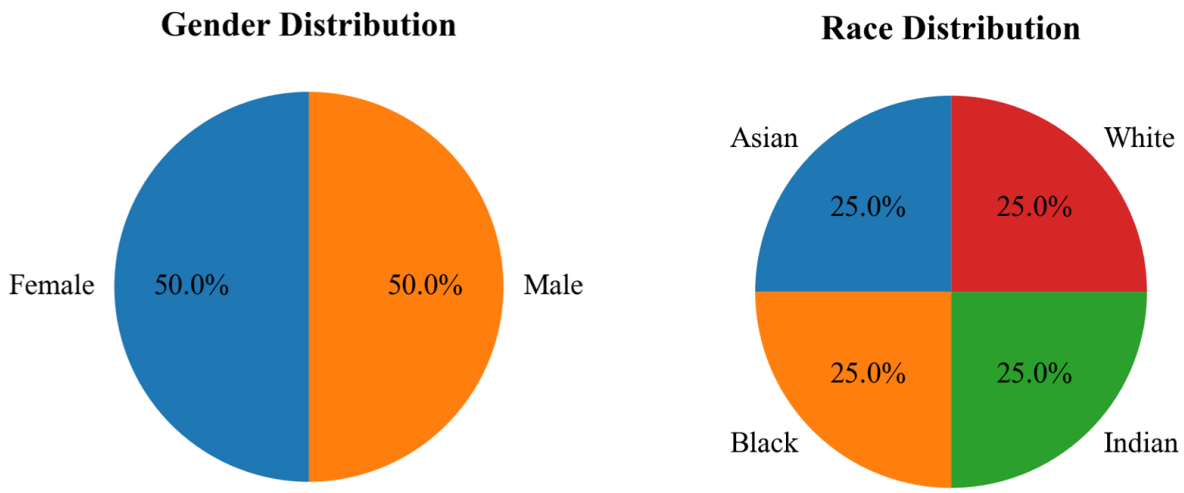

Figure 10: Demographic distribution in the BFW Dataset: Gender (Male/Female) and Ethnicity (Asian, White, Black, Indian) shown as percentages of the total dataset.

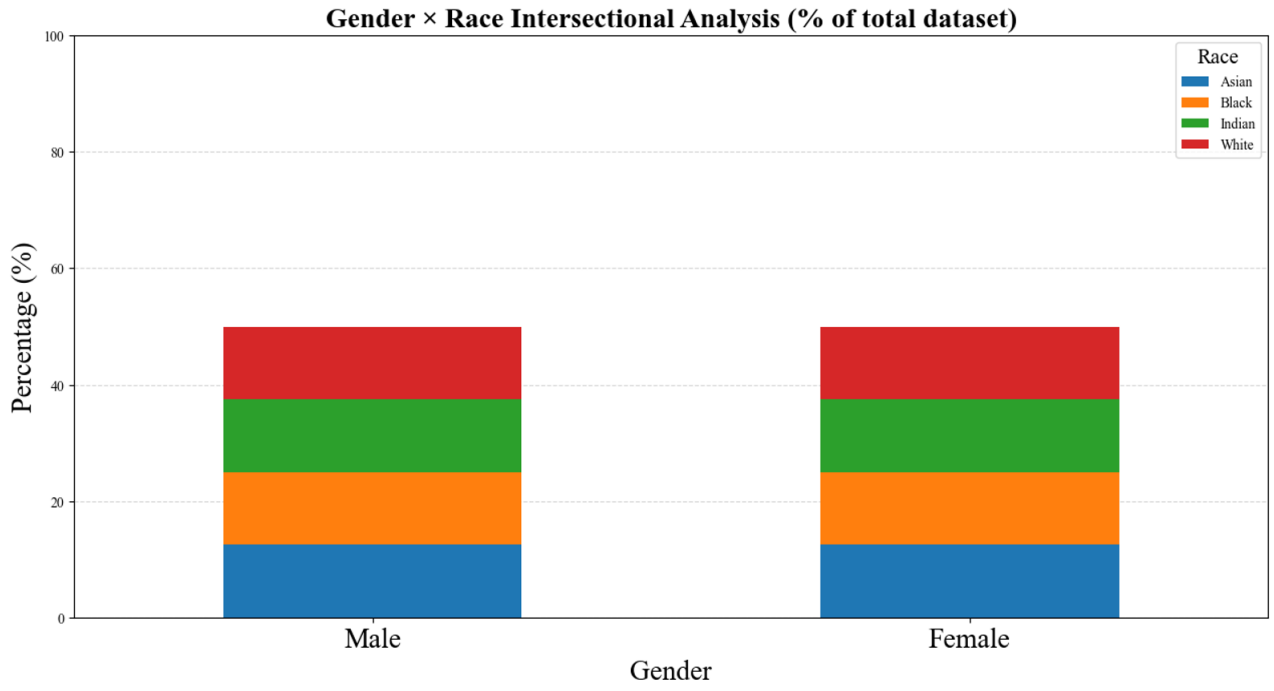

Figure 11: Intersectional analysis of Gender (Male/Female) and Ethnicity (Asian, White, Black, Indian) distributions in the BFW dataset shown as percentages of the total dataset.

6.2 Documentation Analysis Results

The GitHub page directly references two documentation papers written by the original authors of the dataset [48][50]. Likewise, these research papers reference the same github repository as the official source for dataset access and supporting materials. In addition to the references to the papers, the github repository itself also contains information from the documentation papers about the dataset structure, demographic composition and subgroup definitions.

The Balanced Faces in the Wild (BFW) dataset is introduced as a dataset that is balanced across both ethnicity and gender groups. The development of this dataset was motivated by the broader problem of bias in machine learning, particularly in facial recognition systems [50, p. 4]. Robinson et al. explain that many machine learning models inherit bias from the data on which they are trained [50, p. 1].

Robinson et al. provide a simple analogy to explain the concept of bias. If a vehicle detection model is trained primarily on images of trucks, it may become highly effective at recognizing trucks but fail to accurately detect ordinary passenger cars. Similarly, if a facial recognition model is trained mostly on faces from one demographic group, it learns stronger representations for that group and performs less accurately for others. This imbalance in learning leads directly to demographic bias [48, p. 2]

According to the Robinson et al., unlike many other face datasets, where some demographic groups are heavily underrepresented or overrepresented, BFW ensures equal representation across subgroups. The authors define four ethnic categories: Asian, Black, Indian, and White. These are combined with two gender categories: Female and Male, resulting in a total of eight demographic subgroups: Asian Female (AF), Asian Male (AM), Black Female (BF), Black Male (BM), Indian Female (IF), Indian Male (IM), White Female (WF), and White Male (WM). Each subgroup contains 100 subjects, with 25 face images per subject, resulting in a total of 20,000 images and 800 identities. This design ensures balance in the number of faces per subject, individuals per ethnicity, and ethnicities per gender [50, p. 4].

The dataset was constructed using images that are publicly available on the internet. The construction of BFW involved both automated and manual validation procedures. Initially, the researchers used a pre-trained name-ethnicity classifier to generate subject proposals. A name-ethnicity classifier is a machine learning model trained to predict a person's likely ethnic background based on their name. This classifier was used only for initial candidate selection.

After name-based filtering, the corresponding facial images were further refined using separate ethnicity and gender classifiers applied directly to the face images. These image-based classifiers helped verify whether the facial appearance aligned with the proposed subgroup labels. Following automated filtering, four experts in facial recognition manually validated all selected data. First, they verified whether each individual truly belonged to the assigned demographic subgroup. Second, they inspected all facial images to confirm that each image correctly matched the identity of the selected subject. Only subjects and samples unanimously verified by all annotators were retained in the final dataset [50, p. 4, 5].

Robinson et al. emphasize that subgroup definitions in BFW are based on physical facial characteristics most commonly associated with the respective subgroup rather than precise definitions. They acknowledge that race and ethnicity are complex and often imprecisely defined concepts [50, p. 5]. To define subgroups, the researchers referred to categories used by the U.S.

Census Bureau. They openly recognize that such labels are oversimplified and cannot fully represent human identity. However, they say that these simplified categories remain useful for subgroup-level bias analysis in computer vision systems. Robinson et al. explicitly state that BFW is not intended to provide a perfect or final solution to demographic classification but rather to serve as a foundation for future research on fairness and bias mitigation [50, p. 6].

7 Thematic Analysis

7.1 Demographic Imbalance

Analysis of both the MORPH-II dataset and the VGG-Face2 dataset reveal clear imbalances across race, gender, and age, suggesting that certain population groups are represented much more heavily than others. As shown in Figure 3, for the MORPH-II dataset, gender representation is heavily skewed toward male subjects, with 84.4% of the dataset consisting of male images and only 15.6% female images. This imbalance is problematic because models trained on such data are exposed far more frequently to male facial features, allowing stronger feature learning for that group. Female faces, being significantly underrepresented, provide fewer training examples and less variation for the model to learn from. As a result, systems trained on MORPH II may achieve better performance for male subjects while showing weaker recognition accuracy for female subjects. The VGGFace2 is less imbalanced in terms of gender, with a distribution of 59.5% male images and 40.5% female images 7.

However, race and ethnicity distribution is highly uneven for both datasets. For the MORPH-II dataset, analysis shows that 80.1% of the dataset consists of Black subjects, while only 19.9% represents White subjects 3. This indicates a substantial underrepresentation of White subjects within the dataset. Examining the intersection of race and gender reveals additional patterns of imbalance within the MORPH-II dataset. Among Black subjects, even though 86.61% of the images belong to male individuals, only 13.39% represent female individuals 5. A similar pattern is observed among White subjects, where 75.55% of images are male and only 24.45% are female. These results indicate that the dataset is not only imbalanced with respect to race and gender individually, but also across their intersection. In both racial groups, male subjects substantially outnumber female subjects, with the disparity being particularly pronounced among Black individuals. Consequently, Black females represent the least visible demographic group within the dataset, while Black males constitute the most heavily represented group.

The analysis of the VGGFace2 dataset reveals that approximately 79.4% of the images are categorized as Caucasian Latin, while only 8.3% are African American, 6.2% are Asian Indian, and 6.1% are East Asian 7 This demonstrates a strong overrepresentation of Caucasian Latin subjects relative to the other racial groups included in the dataset.

Age representation also reveals imbalance. The largest proportion of the MORPH-II dataset falls within the young adult age group, with 38.68% of individuals aged 18–29 and 29.80% aged 30–39 4. Together, these two categories account for approximately 68.5% of the dataset. In comparison, 23.77% of subjects are between 40–49 years old, while only 7.76% are aged 50 and above. This demonstrates a clear underrepresentation of older individuals, particularly those over

the age of 50.

For the VGGFace2 dataset the younger adult age groups are also the largest, with 35.82% of subjects between 18–29 years old and 33.02% between 30–39 years old 8. Together, these two groups account for nearly 69% of the dataset. In contrast, only 15.99% of subjects are between 40–49 years old, and 15.16% are aged 50 and above. This again shows a underrepresentation of older individuals, while younger adults dominate the dataset.

In contrast to MORPH-II and VGGFace2, the Balanced Faces in the Wild (BFW) dataset contains an equal gender distribution, with 50% male subjects and 50% female subjects 10. Similarly, the ethnicity distribution is balanced across the four demographic groups included in the dataset, with Asian, Black, Indian, and White subjects each representing 25% of the total dataset. The intersectional analysis presented in Figure 11 further demonstrates that this balance is maintained across combinations of gender and ethnicity. Each ethnic group is represented equally within both the male and female categories, resulting in a uniform distribution across all gender-race intersections.

7.2 Consequences of Imbalance

This imbalance in datasets has important implications for facial recognition systems. In the case of VGGFace2, machine learning models trained on such data are exposed much more frequently to the majority group of Caucasian Latin, allowing them to learn stronger and more detailed feature representations for that category. In contrast, other groups, e.g., African American, Asian Indian and East Asian, are underrepresented and receive less exposure during training, which increases the likelihood of lower recognition accuracy and higher error rates [7]. In MORPH-II, the significant underrepresentation of White subjects within the dataset limits the model's ability to generalize effectively to White individuals. This is especially problematic in facial recognition systems, where unequal performance across demographic groups can lead to serious consequences in areas such as identity verification. For example, if a facial recognition system trained on such a dataset is used in law enforcement, underrepresented groups may face higher false positive rates, meaning the system incorrectly identifies an innocent person as a potential match. This can occur simply because the model has had fewer training examples from which to learn accurate distinguishing features for those groups, making its predictions less reliable. This can lead to unjust suspicion or even wrongful arrest. In such cases, technical bias in the dataset moves beyond reduced accuracy and becomes a matter of real social and legal harm [8][14].

As discussed before, not only underrepresentation but also overrepresentation can be problematic in some contexts. In the case of MORPH II, the strong overrepresentation of Black individuals raises important concerns, particularly because of the context in which the data was collected. Since MORPH II is based on mugshot images, the overrepresentation is not a neutral characteristic of the data, but rather a reflection of historical and structural biases in law enforcement practices, including disproportionate arrest rates and surveillance of certain communities [1]. The MORPH II dataset is derived from individuals who entered the criminal justice system rather than from a neutral sample of the general population. As a result, the dataset does not represent society as a whole, but instead represents a highly specific and institutionally shaped subset of the population

[3].

The imbalanced age distribution can also have consequences. In both MORPH-II and VGGFace2, younger faces are more frequently represented and therefore provide more training examples, while older faces receive less representation. As a result, models may learn age-related facial patterns more effectively for younger individuals and perform less accurately on older populations. This can lead to weaker recognition performance for older populations. A dataset that is heavily centered around younger adults may produce systems that generalize poorly across the full age spectrum.
In addition, if the dataset is used for tasks specifically related to age analysis, such as age estimation, the uneven distribution creates further limitations. Since individuals above the age of 40 represent a much smaller portion of the dataset, this can reduce the reliability of age-based predictions and make it more difficult to draw meaningful conclusions about performance across all age categories.

7.3 Reflection in Documentation

The way in which the MORPH II dataset is described in the documentation reveals additional issues. The authors of the paper do not seem to realize the implications of the highly uneven demographic distribution within the dataset, nor do they treat it as a potential issue. Notably, the dataset is even described as representing a "diverse population," despite the reported numbers proving the contrary. This suggests a disconnect between how the dataset is characterized and its actual composition, further reinforcing the lack of critical awareness regarding representational bias [5][37].

Despite how large the imbalance in ethnicities is, the original VGGFace2 paper does not clearly communicate this. The paper repeatedly emphasizes that the dataset contains a "wide range of ethnicities" and presents diversity as one of its main strengths. However, no explicit numerical breakdown of racial distribution is provided. It is remarkable that the limitation of ethnic imbalance is only briefly mentioned once in brackets, where Cao et al. note that ethnic balance is limited by the distribution of celebrities and public figures. This brief acknowledgment does not reflect the scale of imbalance in this dataset, particularly the fact that nearly 80% of the dataset belongs to one single category (e.g. Caucasian Latin). As a result, the documentation may create the impression of balanced diversity, while the actual distribution tells a completely different story [10].

In both the MORPH-II and VGGFace2 documentation, demographic composition is presented primarily as a descriptive characteristic of the dataset rather than as a factor that may influence model performance. There is no discussion provided of how the substantial demographic imbalances within the datasets could affect recognition accuracy across different groups. As a result, the potential implications of demographic imbalance for fairness, bias, and model generalizability remain largely unexplored [5][10][47].

The documentation of the BFW dataset demonstrates a substantially more critical engagement

with demographic bias than the documentation of either MORPH-II or VGGFace2. Robinson et al. explicitly identify bias in facial recognition systems as the primary motivation for creating BFW. Throughout the paper, they recognize that machine learning models can inherit and amplify biases present in training data, and they discuss how unequal representation can lead to disparities in model performance across demographic groups.

The BFW documentation provides a clear and transparent description of the dataset's demographic composition. Robinson et al. explicitly define the demographic subgroups included in the dataset and explain how balance was achieved across ethnicity and gender categories. They acknowledge that race and ethnicity are complex social constructs that cannot be perfectly captured through fixed categories. Rather than presenting their subgroup definitions as objective or comprehensive, they explicitly describe them as simplified categories intended for bias analysis. This recognition of the limitations of demographic labels reflects a more critical and nuanced approach than that found in the documentation of the other datasets [51].

7.3.1 Evaluation of Dataset Documentation

The evaluation of documentation practices across MORPH-II, VGGFace2, and BFW is summarized in Table 2. The table assesses each dataset's documentation using a structured set of criteria derived from common principles in Datasheets for Datasets [19], Dataset Nutrition Labels [28], and Data Cards [45]. These frameworks were analyzed to identify the types of questions they consistently encourage dataset creators to address. Datasheets for Datasets were used to structure criteria related to dataset motivation and intended use, data collection processes, annotation procedures, and documentation of ethical considerations. Data Cards were used to structure criteria related to concerning dataset composition and structure, including detailed breakdowns of dataset features, information about dataset publishers, motivations for dataset creation, and discussion of risks, limitations, and trade-offs associated with the use of the dataset. Dataset Nutrition Labels informed criteria related specifically to information on the origin of the dataset and how the data has been collected.

Each criterion was evaluated and categorized as fully addressed, partially addressed, or not addressed. This distinction was intended to capture not only the presence of documentation but also its level of detail and critical engagement. "Partially addressed" indicates that the dataset documentation includes some relevant information for a given criterion, but that the coverage is incomplete, lacks sufficient detail, or does not explicitly engage with the implications of that aspect. In contrast, "fully addressed" refers to cases where the documentation provides explicit discussion of the criterion, while "not addressed" indicates that the aspect is absent from the documentation.

As discussed in the preceding analysis, the results show clear differences in how datasets document demographic composition and associated limitations. MORPH-II and VGGFace2 primarily provide descriptive information about dataset structure, but offer limited critical reflection. There is limited transparency regarding dataset construction and design choices. Several aspects related to bias, fairness, and downstream risks are either minimally addressed or not discussed at all, despite the presence of substantial demographic skew in both datasets.

In contrast, BFW consistently provides more explicit documentation of demographic categories, dataset limitations, and potential sources of bias, and additionally includes discussion of fairness implications. Table 2 provides an overview of the results from the documentation analysis presented earlier.

Table 2: Documentation evaluation of MORPH-II, VGGFace2, and BFW.

Question:	**MORPH-II**	**VGGFace2**	**BFW**
Dataset Overview			
Is the dataset purpose stated?	🟧	🟩	🟩
Are creators/institutions identified?	🟩	🟩	🟩
Data Collection			
Is the source of the data explained?	🟧	🟩	🟩
Is collection process described?	🟥	🟩	🟩
Is consent/privacy addressed?	🟧	🟥	🟧
Dataset Composition			
Is demographic composition reported?	🟩	🟧	🟩
Are demographic imbalances acknowledged?	🟥	🟧	🟩
Annotation			
Are labels described?	🟩	🟧	🟩
Are annotation procedures documented?	🟩	🟧	🟩
Are subjective labeling decisions discussed?	🟧	🟥	🟩
Bias & Accountability			
Are potential sources of bias discussed?	🟧	🟥	🟩
Are fairness implications discussed?	🟥	🟥	🟩
Are dataset limitations discussed?	🟥	🟧	🟩
Are risks or potential harms discussed?	🟥	🟥	🟩
Is guidance for responsible/intended use provided?	🟥	🟥	🟧

🟩 Fully addressed 🟧 Partially addressed 🟥 Not addressed

7.3.2 Awareness

A recurring observation throughout the analysis is the limited critical awareness reflected in the dataset documentation. In particular, in the documentation of MORPH-II and VGGFace2, the potential consequences receive little or no discussion.

This limited critical engagement can be understood in terms of what is described as privilege hazard [19]. Those who design, collect, and document datasets often occupy positions of relative privilege, which shape what they notice and what remains invisible. Consequently, decisions about which populations are included, how demographic categories are defined, and what limitations are documented may appear technically neutral to dataset creators, while their implications for underrepresented groups remain largely invisible [14].

The analysis also raises broader questions about what is considered worthy of documentation. Both MORPH-II and VGGFace2 devote considerable attention to technical characteristics such as dataset size, image quality, and overall performance, yet provide only limited discussion of representational limitations and their potential consequences [5][10]. This suggests that demographic imbalance was not treated as a methodological limitation of equal importance. Whether this reflects a lack of awareness or the research priorities of the time cannot be determined from the documentation alone. Nevertheless, the omission itself is significant, as failing to acknowledge representational limitations can reinforce the perception that these datasets are broadly representative and suitable for general purpose use.

The contrast with BFW further illustrates this point. Unlike MORPH-II and VGGFace2, the creators of BFW acknowledge existing performance disparities in facial recognition systems, explain the rationale behind balancing demographic groups, and discusses the limitations of demographic categorization itself [51]. This demonstrates a substantially higher level of awareness regarding the relationship between dataset design, representation and algorithmic fairness.

7.3.3 Definition of Diversity

It is notable that documentation papers of MORPH-II and VGGFace2 describe the datasets as representing a "diverse population," despite the massive imbalance in data [5][10]. This appears to rely on the assumption that diversity is achieved simply by the presence of multiple demographic categories, rather than on balanced representation across those categories. This interpretation overlooks the extreme skewness in the actual distribution of the data, where one group overwhelmingly dominates the dataset.

This reflects a broader issue in the way fairness criteria are sometimes reinterpreted in practice. For example, the Equal Employment Opportunity Commission's 4/5ths rule is a guideline used in employment discrimination analysis to identify potential disparate impact [1]. If a protected group is selected at a rate lower than 80% of the rate of the highest-selected group, this may indicate possible discrimination and prompt further investigation. Although this was originally intended as a minimum screening threshold to flag cases for further review, in practice it is sometimes treated as a goal in itself, where meeting the 80% level is taken as evidence that fairness has been achieved [1].

In this sense, diversity is treated as a matter of category inclusion rather than proportional representation. This means that even minimal representation of a group is sometimes presented

as sufficient evidence of diversity, despite significant underlying imbalance. Therefore, the concept of diversity is presented primarily in terms of variety rather than equitable representation. This suggests that the dataset creators do not fully seem to be aware or realize how strongly such imbalance can affect representational fairness and model behavior.

7.4 Sampling Bias

Furthermore, the VGGFace2 dataset relies on images of celebrities and public figures collected from Google Image Search [10]. While this approach provides a large collection of facial images, it introduces a form of sampling bias because the dataset inherently reflects patterns of visibility in online media rather than the demographic composition of the global population. Celebrities and public figures might differ from the general population in terms of appearance.

Also, since the dataset is shaped by media representation and search engine retrieval processes, it may reinforce existing societal and cultural biases. Individuals that are already more visible are more likely to be included in the dataset, while less visible groups are further marginalized.

Patterns of visibility are themselves influenced by broader social and historical inequalities. Historical inequalities related to gender, race, colonialism, and colorism have shaped representation in media [7]. Skin tone and racial representation provide a clear illustration of this issue. Buolamwini notes that despite efforts to create globally diverse facial recognition datasets, some benchmark datasets remained overwhelmingly composed of lighter-skinned individuals [7]. This is particularly striking given research by Nina Jablonski demonstrating that the majority of the world's population would be classified toward the darker end of common skin tone scales [32]. The overrepresentation of lighter-skinned individuals therefore cannot be explained simply by global population demographics.

Colorism contributes to these patterns. Unlike racism, which distinguishes between racial groups, colorism operates within and across racial groups by assigning greater social value to individuals with lighter skin tones. For example, even when dataset creators aim to collect facial images representing specific demographic groups, such as Asians, broader patterns of media representation can still shape who is most visible. Across many Asian contexts, despite the vast diversity of skin tones within populations, entertainment and beauty industries have historically elevated lighter-skinned actors and actresses [7][30].

As a result, they are more likely to appear in publicly available image collections that are later used to construct facial recognition datasets. Consequently, these datasets may systematically overrepresent groups that already possess greater social visibility, while leaving out or underrepresenting other groups [8].

7.5 Labels

Another central issue identified in this analysis is the lack of accessibility and organization of labels that come with the dataset. For the MORPH-II and VGGFace2 dataset, the labels must be sourced from multiple external repositories, rather than as a single original dataset release with everything included. This creates some challenges as it requires additional effort to locate all the different labels. Besides that, it also makes intersectional analysis impossible for the VGGFace2

dataset as the labels with demographic attributes are fragmented and not available in a single unified dataset.

In addition, the documentation of label construction is insufficient, especially for the VGGFace2 dataset. In the VGGFace2 documentation paper, there is no clear information about how the demographic labels were obtained. This shows a lack of transparency and raises concerns about label reliability, reproducability and possible annotation biases.

An issue is that even when datasets might be initially released with annotation files, their availability is often not guaranteed over time [44]. Although many datasets are released with the intention of supporting open scientific research, it has been noted by Scheuerman et al. that once datasets are made publicly available, it becomes extremely difficult to maintain control over their distribution and usage [52]. Many datasets are distributed through simple URLs rather than persistent repositories, meaning that both the image data and their associated label files may become inaccessible as links expire or hosting platforms change [52]. This lack of persistence creates a fragmented research environment in which datasets may still be actively used by researchers who previously downloaded them, while becoming partially or fully unavailable to new users. Consequently, access to annotations becomes unevenly distributed across the research community, depending on when the dataset was obtained. This issue is particularly evident in datasets such as MORPH-II and VGGFace2, where demographic labels are not consistently included and not released as one original publication with the data and documentation.

7.6 Demographic Categorization

Race and ethnicity are frequently used demographic variables in facial recognition datasets. However, despite being treated as straightforward dataset attributes, race and ethnicity are not objective or universally defined characteristics. Rather, they are social categories whose meanings vary across cultures, countries, and time. Race and ethnicity are not always clearly distinguished, and their definitions may vary depending on institutional or national standards. Race is often associated with perceived physical characteristics, while ethnicity is more closely related to cultural identity, including language, ancestry and shared heritage [39].

When race and ethnicity are incorporated into datasets, complex identities must be translated into a limited number of categories. Individuals who identify with multiple ethnicities or whose identities do not fit neatly within predefined categories are often forced into a single classification [55].

The way these categories are assigned also differs across datasets. In many cases, they are assigned by researchers or annotators based solely on an individual's physical appearance, for the purpose of dataset annotation. This approach is particularly problematic because it assumes that racial or ethnic identity can be reliably inferred from visual characteristics. As a result, these classifications often reflect the perceptions and assumptions of annotators rather than individuals' own identities, introducing subjective judgements into what are subsequently treated as objective demographic labels [55].

The comparison between the three datasets demonstrates that demographic categorization is itself a design decision rather than an objective property of the data. The MORPH-II dataset mention 'race' labels, while the VGGFace2 and BFW mention 'ethnicity' labels. Yet none of them

fully explains why these concepts were selected. This shows a lack of methodological transparency.

It is notable that the available labels used for the analysis of MORPH-II are restricted to Black and White categories, which is not representative of population diversity in the real world. This categorization reduces racial identity to a binary distinction, excluding a wide range of other groups.

Another issue with categorization is that gender representation in the datasets is reduced to a binary classification, consisting only of "male" and "female." Within this framing, individuals who do not conform to binary gender expressions, or whose appearance does not align with stereotypical expectations of masculinity or femininity, are therefore either misclassified or excluded from meaningful representation altogether [52][53].

7.7 Accuracy

Accuracy is a central and widely emphasized metric in the evaluation of face recognition systems. In both MORPH-II and VGGFace2, the accompanying documentation primarily focuses on overall accuracy. However, although many face recognition systems often demonstrate high overall accuracy, these aggregate performance measures can conceal significant disparities across demographic groups. Publicly available facial datasets are frequently imbalanced. As a result, the reported performance of face recognition systems is heavily influenced by the majority groups present in the dataset [8].

Because these dominant groups contribute the largest proportion of samples, they disproportionately determine the overall accuracy score. This imbalance creates a serious fairness concern: a model may achieve excellent global accuracy while simultaneously performing poorly on the minority of underrepresented populations [8]. In such cases, the system appears reliable when evaluated on average, but this average masks unequal performance across different demographic categories [1][38].

For example, consider a dataset with 10,000 images where 99% belong to one demographic group and only 1% to another. If a model achieves 99.9% accuracy on the majority group but 0% on the minority group, it still attains an overall accuracy of about 99%. This shows how a high aggregate accuracy can be driven almost entirely by the majority group, while completely masking poor performance on underrepresented populations.

This means that performance ratings are often misleading interpretations of model quality. Consequently, overall accuracy alone is insufficient as a measure of fairness or robustness in face recognition systems, as it can mask systematic disparities affecting underrepresented groups.

8 Reflections on Data Science Education

This section is included to situate Data Science and Artificial Intelligence students within the broader discussion on dataset bias and documentation.

Initially, I intended to conduct an empirical study with students from the university in order to explore how future data scientists perceive dataset bias, documentation practices, and representational issues in practice. However, due to the scope of this thesis, such an experiment was not feasible.

Instead, my own educational experience, together with reflections from a few fellow students, is used to illustrate how these issues are currently encountered within data science training.

These reflections on data science education can offer insight into how dataset bias and documentation are typically introduced and understood in technical training, and why certain forms of bias and representational concerns may remain underexplored.

As a third-year Data Science and Artificial Intelligence student (2023-2026), my experience is that in this field, datasets are often treated primarily as technical resources rather than as objects that require critical evaluation. When working on machine learning assignments and projects, the focus is on achieving overall good model performance. Questions concerning demographic representation, dataset bias, or the social implications of dataset design are rarely discussed.

8.1 My Experience

From my experience, awareness of dataset bias among students remains general. I do not feel that students receive enough training on dataset bias and representation. I had not really thought about the importance of diverse representation in datasets in any depth. While it was generally acknowledged that biased data can lead to biased models, the topic is not emphasized enough or explored in sufficient depth and I did not actively reflect on how representation within datasets shapes model behavior. There is little attention paid to how bias arises through decisions about data collection, sampling, annotation, demographic categorization, or representation, or how these choices influence the behavior of machine learning systems. Dataset bias is often acknowledged as a general concern without being systematically investigated when selecting or using data. As a result, I largely viewed datasets as objective inputs rather than as products of numerous human decisions and assumptions.

This perspective changed when I voluntarily chose to follow the *AI & Society* minor alongside my major. In contrast to the primarily technical focus of my degree program, the minor introduced ethical and societal perspectives on artificial intelligence and encouraged a more critical examination of the data underlying machine learning systems. It was through this minor that I first encountered concepts such as representational bias. I found it remarkable that these topics were introduced through a voluntary minor rather than as a core component of my degree program. For example, in core courses such as *Databases* and *Machine Learning*, I would expect issues such as dataset bias, representational imbalance, and data documentation to be addressed in a more explicit and integrated manner. However, in practice, in the Databases course, we learned how to design databases, store, modify, and retrieve information, as well as how databases function and how to optimize data retrieval. However, there was little discussion of datasets as they are used in machine learning, nor of their broader role in shaping model behavior and outcomes. In addition, the machine learning course only briefly mentioned bias generally, without engaging in any detailed coverage of representation issues.

What is striking about this omission is that it places critical issues related to dataset bias and representation outside the main technical training of future data scientists and AI practitioners, despite their direct influence on model behavior and decision outcomes. In particular, before I started working on this thesis, I had never heard of dataset documentation frameworks such as

Datasheets for Datasets, Dataset Nutrition Labels, or Data Cards. I was unaware that formal frameworks even existed to document how datasets were collected, annotated, and intended to be used.

This lack of awareness is significant, as it suggests that without familiarity with such documentation practices, I would be unlikely to rely on documentation of datasets to inform me about a dataset's limitations, assumptions, and appropriate use cases, even in future professional practice. Moreover, if I am not aware that dataset bias is something that can exist in the first place, it is unlikely that I would actively look for it or think to question it when working with a dataset, simply because it would not be part of what I consider relevant at that point.

Only in the third year of my bachelor's program did I follow a mandatory course called *AI & Ethics*. Although this course addressed important topics related to the societal implications of artificial intelligence, it was limited in scope, 3 EC, equivalent to half a semester. As a result, the treatment of ethical issues remained introductory. In particular, the course tended to emphasize broad, future-oriented questions about artificial intelligence, including hypothetical scenarios and possible long-term societal risks. While these perspectives are also valuable, less attention was given to already existing issues in deployed machine learning systems, such as dataset bias, representational imbalance, and unequal performance across demographic groups. As a result, ethical reflection was often framed around "big-picture" or speculative concerns about the future of AI, rather than the more immediate and empirically observable issues present in current systems. I think emphasis on immediate concerns would better support the development of practical skills needed to evaluate data critically and to make informed decisions in real-world machine learning contexts.

8.2 Shared Experience Among Students

These observations are not unique to my own experience but also align with perspectives provided by other students in the field of Data Science and Artificial Intelligence, who were asked to write down their views on these topics.

A student currently doing a master in Data Science & AI stated:

> "As a Master student in Data science and Artificial Intelligence, I have had very little education on the topic of bias in datasets and dataset documentation. My first experience with dataset bias was when I attempted to find a dataset for a computer vision project I was working on for my master. The topic I was tackling was the degradation of Computer Vision model performance for darker skin. I, unfortunately, could not carry through with this topic as I could not find a suitable dataset that had enough diversity with regards to dark skin. Even so-called diverse datasets mostly consisted of lighter skin brown/black people. This made me realize the enormous bias that is present in a lot of datasets, and how it will drastically affect real-life results if not reflected upon prior to usage.
>
> — Master student in Data Science and Artificial Intelligence, TU Delft

This experience suggest that exposure to dataset bias often emerges through practical challenges encountered during project work or personal encounters, rather than through formal education.

A similar pattern is reflected in the perspective of a Bachelor's student in Data Science and Artificial Intelligence, who described a strong assumption of objectivity in data during their studies:

> During my bachelor's in Data Science and Artificial Intelligence, I have never stopped to consider that data is not entirely objective. To me, data appears as objective numbers, similar to mathematics, where subjectivity plays no role at all. When working with data, I do not reflect on how it was collected and what the implications are. I also did not consider that documentation can shape how data is interpreted, and I have never been exposed to documentation frameworks until now.
>
> — Bachelor student in Data Science and Artificial Intelligence, Leiden University

This lack of exposure to documentation practices is further reinforced by another Bachelor's student, who highlighted an absence of awareness regarding dataset documentation:

> As a Data Science and Artificial Intelligence student, I have never thought about any documentation beyond the data itself. I had never heard of structured documentation frameworks such as Datasheets for Datasets or Data Nutrition Labels, and I was not aware that such frameworks even existed.
>
> — Bachelor student in Data Science and Artificial Intelligence, Leiden University

As discussed before, this lack of awareness is significant, as it suggests that these students are unlikely to rely on documentation of datasets, since they do not even know it exists.

Looking back, these findings are surprising to me because datasets form the foundation of every machine learning model. Considerable attention is devoted to selecting algorithms, tuning hyperparameters, and evaluating performance, whereas comparatively little attention is paid to critically assessing the quality, representativeness, and documentation of the data itself. I believe that greater emphasis on these topics within data science education would better prepare students to recognize potential sources of bias and to make more informed decisions when using datasets and building models.

8.3 Possible Experiment Questions

The following questions were initially designed for an empirical study to explore the understanding and practices of a larger group of students, in relation to dataset bias, representational issues, and

dataset documentation in machine learning. Even though the experiment was not within the scope of this thesis, the questions are still included as they illustrate the intended research direction and provide insight into how the topic could be empirically investigated in future work. The questions are organized into three thematic areas: (1) perceptions of bias in datasets, (2) engagement with dataset documentation, and (3) educational preparedness and professional practice.

Perceptions of dataset bias

These questions focus on how dataset bias is understood conceptually, including its perceived causes, significance, and impact on machine learning systems. It also examines awareness of bias during model development. These questions are intended to capture whether bias is recognized as a structural issue within datasets or primarily as a general or abstract concern.

- What do you see as the main causes of bias in machine learning datasets?
- To what extent do you think demographic representation in datasets influences the fairness and performance of machine learning systems?
- How aware do you think people are of dataset bias when developing machine learning models?

Dataset documentation practices

This investigates how dataset documentation is used and understood in practice. It addresses whether documentation is consulted, what information is considered necessary, and how useful structured documentation would be for dataset selection and responsible use. It also includes awareness of established documentation frameworks such as Datasheets for Datasets, Dataset Nutrition Labels, and Data Cards. These questions aim to assess whether documentation is seen as an integral part of the machine learning workflow or as supplementary information.

- How important do you consider dataset documentation in machine learning research?
- What information do you consider essential to include in dataset documentation before a dataset can be responsibly used for research or model development?
- In your experience, how often do people actually consult dataset documentation before using a dataset?
- When using a dataset, would you personally first read its documentation?
- Have you heard of different types of dataset documentation frameworks (Datasheets for Datasets, Dataset Nutrition Labels, Data Cards)? If yes, what do you know about them?
- Do you think more detailed and transparent documentation about limitations would influence dataset selection?

Education and professional practice

This focuses on perceived preparedness to evaluate dataset bias and representation, as well as attitudes towards responsibility and documentation requirements in professional settings.

- Do you feel you have enough training to evaluate dataset bias and representational issues? Why?
- As a dataset consumer, would mandatory dataset documentation requirements make your workflow easier, harder, or unchanged? Why? And what if you were a dataset creator?

9 Conclusions and Further Research

This thesis has examined representational bias in facial recognition datasets and how such biases are reflected in dataset documentation. As shown in the analysis section, the MORPH-II and VGGFace2 datasets contain systematic imbalances in demographic representation. The analysis further shows that there is a gap between the presence of bias in the data and the extent to which it is communicated in documentation, which limits transparency and makes it harder to evaluate dataset suitability for different applications.

Taken together, these findings raise a broader question for the field: when representational biases are present but insufficiently made visible, what changes are required in order to ensure that such biases are identified and addressed rather than overlooked?

A further direction concerns the development of more consistent standards for reporting dataset composition and biases. The analysis in this thesis suggests that current documentation practices are uneven and often incomplete, largely because there are no widely agreed standards for what should be reported and because there is little practical incentive for dataset creators to document these aspects in detail. As a result, there is a need for clearer and more widely adopted reporting standards that define a minimal set of information on dataset representation and biases that should be included by default.

In addition, further empirical work investigating how documentation frameworks lead to measurable improvements in awareness, decision-making, or bias mitigation in practice would be valuable. Measurable results could provide strong motivation for wider adoption of documentation practices and encourage practitioners and researchers to incorporate such frameworks more consistently in the field.

The findings also have implications for data science education. In particular, they suggest that current training may not sufficiently address issues related to dataset bias, representational imbalance, and documentation practices.

In the future, a study with a large group of student participants would be valuable in order to extend the observations and reflections made in Section 8. Particularly to assess whether similar gaps in awareness and practice are present across a broader population of data science students. An empirical study based on the suggested questions in Section 8 would allow for an examination of whether the issues identified in this thesis also stem from limitations in how representational bias and dataset evaluation are currently taught and understood within data science education.

References

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, Barcelona Spain, January 2020. ACM.

[2] J. E. Alderman et al. Tackling algorithmic bias and promoting transparency in health datasets: the standing together consensus recommendations. *The Lancet Digital Health*, 7:e64–e88, 2025.

[3] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *SSRN Electronic Journal*, 2016.

[4] BBC News. Google apologises for photos app's racist blunder. July 2015.

[5] Garrett Bingham, Benjamin Yip, Matthew Ferguson, and Christine Nansalo. Morph-ii: Inconsistencies and cleaning whitepaper. Technical report, NSF-REU Site, University of North Carolina Wilmington, 2017.

[6] danah boyd and Kate Crawford. Critical questions for big data. *Information, Communication & Society*, 15(5):662–679, 2012.

[7] Joy Buolamwini. *Unmasking AI: my mission to protect what is human in a world of machines*. Random House, New York, 2023.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of Machine Learning Research*, volume 81 of *Proceedings of Machine Learning Research*, pages 1–15. PMLR, 2018.

[9] Ben Burgess, Avi Ginsberg, Edward W. Felten, and Shaanan Cohney. Watching the watchers: Bias and vulnerability in remote proctoring software. 2022.

[10] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, Xi'an, China, 2018. IEEE.

[11] Kasia S. Chmielinski et al. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. 2022.

[12] Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, April 2021.

[13] Caroline Criado Perez. *Invisible Women: Data Bias in a World Designed for Men*. Abrams Press, New York, 2019.

[14] Catherine D'Ignazio and Lauren F. Klein. *Data Feminism*. Strong Ideas. The MIT Press, Cambridge, 2020.

[15] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, 1996.

[16] Song Fu, Haibo He, and Zeng-Guang Hou. Learning race from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2483–2509, 2014.

[17] Seeta Pěna Gangadharan and Jědrzej Niklas. Decentering technology in discourse on discrimination. *Information, Communication & Society*, 22(7):882–899, 2019.

[18] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The perpetual line-up: Unregulated police face recognition in america. Technical report, Georgetown Law Center on Privacy & Technology, Washington, DC, 2016.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, December 2021.

[20] Antonio Greco, Gennaro Percannella, Mario Vento, and Vincenzo Vigilante. Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*, 31(7-8):67, November 2020.

[21] Antonio Greco, Alessia Saggese, Mario Vento, and Vincenzo Vigilante. Effective training of convolutional neural networks for age estimation based on knowledge distillation. *Neural Computing and Applications*, 34(24):21449–21464, December 2022.

[22] Greco, Antonio and Percannella, Gennaro and Vento, Mario and Vigilante, Vincenzo. Vggface2 mivia ethnicity recognition dataset. https://mivia.unisa.it/ethnicity-recognition-dataset/, 2020.

[23] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[24] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. Technical Report NIST IR 8280, National Institute of Standards and Technology, 2019.

[25] Guodong Guo and Guowang Mu. Human age estimation: What is the influence across race and gender? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 71–78. IEEE, 2010.

[26] Guodong Guo and Guowang Mu. A study of large-scale ethnicity estimation with gender and age variations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 79–86. IEEE, 2010.

[27] A. Heinke, L. Huang, K. U. Simpkins, et al. Dataset documentation for responsible ai: analysis of suitability and usage for health datasets. *npj Digital Medicine*, 2026.

[28] Sarah Holland, Ahmed Hosny, Sorelle Newman, Jonathan Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[29] Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

[30] Margaret Hunter. The persistent problem of colorism: Skin tone, status, and inequality. *Sociology Compass*, 1(1):237–254, 2007.

[31] Ben Hutchinson, Andrew Smart, Alex Hanna, Remi Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, Virtual Event Canada, March 2021. ACM.

[32] Nina G. Jablonski. The evolution of human skin and skin color. *Annual Review of Anthropology*, 33:585–623, 2004.

[33] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, pages 306–316, New York, NY, USA, 2020. Association for Computing Machinery.

[34] Yakhyo Khujayev. Vggface2 112x112. https://www.kaggle.com/datasets/yakhyokhuja/vggface2-112x112, 2023. Aligned and Cropped Dataset.

[35] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1–12, 2014.

[36] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[37] Troy P. Kling. Morph-ii: Feature vector documentation. Technical report, NSF-REU Site at UNC Wilmington, 2017. Technical documentation for the MORPH-II facial image dataset.

[38] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[39] Catherine Lewis, Philip R. Cohen, Devyani Bahl, Elliot M. Levine, and Waseem Khaliq. Race and ethnic categories: A brief review of global terms and nomenclature. *Cureus*, 11(5):e4858, 2019.

[40] Yiming Lin. morph2-protocols: Facial age estimation protocols for the morph2 dataset. https://github.com/yiminglin-ai/morph2-protocols, 2021.

[41] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. FP-Age: Leveraging Face Parsing Attention for Facial Age Estimation in the Wild. *IEEE Transactions on Image Processing*, 34:4767–4777, 2025.

[42] Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda F. Gürses. Questioning the assumptions behind fairness solutions. *arXiv preprint arXiv:1811.11293*, 2018.

[43] Chirag Sai Panuganti. Morph-2. https://www.kaggle.com/datasets/chiragsaipanuganti/morph, 2023. Cropped, aligned, and pre-processed version of the MORPH-II dataset.

[44] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, November 2021.

[45] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery.

[46] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. "you can't sit with us": Exclusionary pedagogy in ai ethics education. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[47] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006.

[48] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.

[49] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Source code and benchmarks for balanced faces in the wild. https://github.com/visionjo/facerec-bias-bfw, 2020.

[50] Joseph P. Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing biases and preserving privacy on balanced faces in the wild. *arXiv preprint arXiv:2103.09118*, 2021.

[51] Joseph P. Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing Biases and Preserving Privacy on Balanced Faces in the Wild. *IEEE Transactions on Image Processing*, 32:4365–4377, 2023.

[52] Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, October 2021.

[53] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis and image labeling services. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–33, 2020.

[54] Harini Suresh and John V. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*, pages 1–9, New York, NY, USA, 2021. Association for Computing Machinery.

[55] Gerardus Adrianus van Schie. *The Datafication of Race-Ethnicity: An Investigation into Technologically Mediated Racialization in Dutch Governmental Data Systems and Infrastructures*. Phd dissertation, Utrecht University, Utrecht, The Netherlands, 2022.

[56] Visual Geometry Group, University of Oxford. Vggface2 dataset for face recognition. https://github.com/ox-vgg/vgg$_f ace2$, 2018.

[57] Yunqian Wen, Li Song, Bo Liu, Ming Ding, and Rong Xie. Identitydp: Differential private identification protection for face images. *IEEE Transactions on Information Forensics and Security*, 2023.

[58] Genta Indra Winata, David Anugraha, Emmy Liu, Alham Fikri Aji, Shou-Yi Hung, Aditya Parashar, Patrick Amadeus Irawan, Ruochen Zhang, Zheng-Xin Yong, Jan Christian Blaise Cruz, Niklas Muennighoff, Seungone Kim, Hanyang Zhao, Sudipta Kar, Kezia Erina Suryoraharjo, M. Farid Adilazuarda, En-Shiun Annie Lee, Ayu Purwarianti, Derry Tanti Wijaya, and Monojit Choudhury. Datasheets aren't enough: Datarubrics for automated quality metrics and accountability. *arXiv preprint arXiv:2506.01789*, 2025.

[59] Benjamin Yip, Garrett Bingham, Katherine Kempfert, Jonathan Fabish, Troy Kling, Cuixian Chen, and Yishi Wang. Preliminary studies on a large face database. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2572–2579, Seattle, WA, USA, December 2018. IEEE.