Fairer Datasets for Advancing Responsible AI Systems

Siddharth Jaiswal

Indian Institute of Technology Kharagpur, India
siddsjaiswal@kgpian.iitkgp.ac.in

Research Direction

Modern AI systems are data-guzzling machines, heavily reliant on large-scale multimodal data sourced from the digital and physical world. Examples of such systems are chatbots like OpenAI's ChatGPT and image/video generative models like Stability AI's Stable Diffusion. These platforms have brought about a paradigm shift in how technology is perceived, used and abused in human society. While the benefits of these systems cannot be understated, they have also been observed to reproduce unprecedented discriminatory and biased behaviour against minority and marginalized members of society. For example, in my previous studies, I have observed large-scale consistent biases on Face Recognition Systems against dark-skinned people (Jaiswal et al. 2022; Jaiswal, Verma, and Mukherjee 2024) & people from the Global South (Jaiswal et al. 2024) and, on other AI systems against non-binary gender groups (Jaiswal and Mukherjee 2022; Jaiswal, Verma, and Mukherjee 2023).

One of our recent projects involved studying the relationship between datasets, architectures, and loss functions, and how these factors impact the accuracy and disparity in Face Recognition Systems (Jaiswal et al. 2025). Two striking observations were– (a) the same dataset reporting biases against opposite binary gender groups in two different FRS models, (b) two different datasets reporting biases against opposite binary gender groups on the same FRS model. Thus, it is apparent that datasets play a very important role in the training and testing of such AI models.

Existing benchmark and training datasets are prepared from a very narrow, Euro-centric perspective and thus represent a very small, albeit privileged, minority of the world. AI models, when trained on such datasets, perpetuate discriminatory biases against the Global Majority population, composed of disadvantaged and underprivileged individuals.

Thus, it is highly critical to prepare fairer, more diverse, and more representative datasets that can be used to benchmark, audit, and train AI models. Such datasets will not only reduce existing biases but also mitigate the potential for future harm, especially for groups belonging to countries where legal recourse is not yet available.

Problem Statement: The primary aim of my research is to develop *diverse and fair datasets*, with a specific focus on marginalized demographics like the Global South and/or non-binary gender groups. I am working towards creating two types of datasets– (a) Real (Jaiswal et al. 2024) and Synthetic (under submission) face datasets from Global South for auditing/training Face Recognition Systems. These are used in tasks like face detection and facial attribute analysis, and (b) Datasets with non-binary gender labels, across modalities like image (Jaiswal and Mukherjee 2022) and text (Jaiswal, Verma, and Mukherjee 2023) for more inclusive retrieval (visual search enabled e-commerce) and classification (text-based gender prediction) tasks.

Research Questions and Contributions

I am addressing the following research questions—

RQ1. Face datasets from the Global South: Face Recognition Systems are trained on large-scale face image datasets for tasks like face detection and facial attribute analysis. These datasets are curated from various sources on the internet, primarily from the Western hemisphere and Caucasian ethnicity. Thus, they underperform for the Global Majority ethnicities and geographies. There is a need to curate benchmark datasets from these regions, not only to audit existing platforms, but also to train/fine-tune them for fairer performance.

Gaps Identified: Existing face datasets, while large-scale in terms of size, suffer from three major shortcomings— (i) lack of representation from the Global South demographics, (ii) usage of images of regular individuals without consent and (iii) lack of non-standard adversarial images. Thus, the models trained on these datasets report discriminatory biases, especially against individuals from the Global South and people of color.

Proposed Solution: I have attempted to address all three issues described above through the following means. To address Gap 1, I have curated two types of face datasets— (a) Real face datasets with more than 50% images from the Global South countries. One of the datasets, FAR-Face (Jaiswal et al. 2024) has proved to be highly adversarial for existing FRSs and helped expose biases in advanced models like Vision Transformers and VLMs (Jaiswal et al. 2025). I have also expanded this dataset to be more geographically and culturally representative. (b) Synthetic face datasets curated using text-conditioned diffusion models that

use an aligned VLM to generate realistic and detailed input prompts. My quantitative experimental results, along with a human survey, have demonstrated that the dataset is highly realistic and diverse. My pipeline is highly scalable and only requires a small set of seed images from the demographic that needs to be generated.

The real datasets are composed of public-facing individuals such as sportspersons and politicians, and the synthetic dataset is privacy-respecting (I have experimentally verified the face matching accuracy against the seed set and observed less than 40% accuracy). This addresses Gap 2.

To address Gap 3, I have additionally curated adversarial variants of all face images in the real face datasets. These adversarial variants simulate simple, real-world challenges like face masks, grainy film, liquid on camera lenses, etc. I am working on developing strategies to automate this process for synthetic face images.

RQ2. Datasets for non-binary inclusion: A majority of AI tasks that either use gender as a feature or for prediction only focus on binary gender labels, ignoring non-binary identifying individuals. This can be attributed to a lack of societal understanding of non-binary labels, developer oversight and/or legal restrictions. These individuals already face real-world discrimination in society, and digital discrimination (Jaiswal and Mukherjee 2022; Jaiswal, Verma, and Mukherjee 2023) only compounds their challenges.

Gaps Identified: Existing gender based datasets rarely include non-binary labels. While AI applications are already biased against the female gender for a variety of tasks, due to a lack of existing labels for the non-binary genders, the biases multiply. Hence, there are two major challenges that need addressing– (i) Lack of gender-inclusive datasets and, (ii) Lack of qualitative studies exploring gender-inclusive design of AI platforms.

Proposed Solution: I have attempted to address the two issues described above through the following means. To address Gap 1, I have curated two gender-inclusive datasets with more than 50% data points from the non-binary gender group. The first dataset is a fashion image dataset used for auditing visual search based e-commerce platforms for clothing recommendations (Jaiswal and Mukherjee 2022). This manually curated dataset helped expose how Amazon's visual search feature treats non-binary image inputs. Next, I curated a textual comment dataset (Jaiswal, Verma, and Mukherjee 2023) with more than 1M data points from Reddit & Tumblr with self-annotated gender labels. This dataset has proven useful in auditing text-based gender classifiers and has exposed how these models tend to classify non-binary users as feminine.

To address Gap 2, I am currently working on developing a study to audit non-binary users and identify the challenges they face while using AI-based computing systems, as well as how existing design decisions are discriminatory and biased. This will allow not only to understand the shortcomings of existing platforms from a user-centric perspective but also to design more effective and inclusive systems.

Current Focus & Future Directions

(A) I have developed a real (Jaiswal et al. 2024) and a synthetic face dataset (under submission), specifically focusing on a subset of Global South countries. Presently, I am working towards increasing the size and geographical coverage of these datasets. This will allow developers & researchers to perform more fine-grained evaluations as well train more context-appropriate models. This will help me address Gaps 1 and 2 in RQ1.

I am also working towards designing automated strategies to create realistic adversarial variants of synthetic face images for audits under real-world conditions. This will help me address Gap 3 in RQ1.

(B) I plan to extend my research in curating non-binary inclusive datasets (Jaiswal and Mukherjee 2022; Jaiswal, Verma, and Mukherjee 2023) in two major directions– (i) other application domains in generative AI models like VLMs and T2I models and (to expand upon the Gap 1 identified in RQ2), (ii) qualitative studies with non-binary participants to understand their lived experiences and the discrimination they face while interfacing with modern AI tools (current focus, to address Gap 2 in RQ 2). This will help motivate better design decisions at both the dataset preparation as well as model deployment stages.

References

Jaiswal, S.; Basu, S.; Sikdar, S.; and Mukherjee, A. 2025. Exploring Disparity-Accuracy Trade-offs in Face Recognition Systems: The Role of Datasets, Architectures, and Loss Functions. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Jaiswal, S.; Duggirala, K.; Dash, A.; and Mukherjee, A. 2022. Two-face: Adversarial audit of commercial face recognition systems. In *AAAI ICWSM*.

Jaiswal, S.; Ganai, A.; Dash, A.; Ghosh, S.; and Mukherjee, A. 2024. Breaking the global north stereotype: A global south-centric benchmark dataset for auditing and mitigating biases in facial recognition systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Jaiswal, S.; and Mukherjee, A. 2022. Marching with the Pink Parade: Evaluating Visual Search Recommendations for Non-binary Clothing Items. In *ACM CHI EA*.

Jaiswal, S.; Verma, A. K.; and Mukherjee, A. 2023. Auditing gender analyzers on text data. In *IEEE/ACM ASONAM*.

Jaiswal, S. D.; Verma, A. K.; and Mukherjee, A. 2024. Mask-up: Investigating Biases in Face Re-identification for Masked Faces. *arXiv preprint arXiv:2402.13771*.