IT UNIVERSITY OF COPENHAGEN

Data In Wild - Project Report

Investigating the Impact of Demographic Diversity on Model Generalization in Facial Expression Recognition (FER) Systems

Gergo Gyori (gegy@itu.dk)
Katalin Literati-Dobos (klit@itu.dk)
Ivan Petrov (ivpe@itu.dk)
Marcin Sroka (msro@itu.dk)

2025-01-03

Abstract

Facial Expression Recognition (FER) systems are crucial for applications in healthcare, human-computer interaction, and security. However, the effectiveness and fairness of these systems are often limited by demographic biases in their training datasets. This study focuses on two objectives: first, analyzing trends in FER research by creating a reusable dataset of 115 academic papers, including metadata and full texts, to explore frequently cited FER datasets and their reported metrics. Second, we evaluate demographic diversity in four prominent FER datasets using automated and manual annotation methods for attributes such as age, gender, and ethnicity. Our findings reveal that datasets with limited demographic diversity show higher accuracy but reduced generalizability, while more diverse datasets achieve broader real-world applicability at the cost of lower accuracy. By highlighting these trade-offs and demonstrating the utility of demographic annotations, this project emphasizes the importance of constructing equitable FER datasets and models.

1 Introduction

Facial Expression Recognition (FER) systems are widely used in fields such as healthcare, human-computer interaction, and security, where accurately interpreting human emotions has significant implications. These systems rely heavily on the datasets used for training, and the quality and diversity of these datasets significantly impact their performance and fairness. Imbalances in factors such as gender, ethnicity, and age group representation can result in models that perform inconsistently across different groups, raising concerns about fairness and generalization [2, 11].

While FER research continues to grow, many commonly used datasets face challenges related to demographic representation. The Affective Faces Database (AffectNet) [13] is one of the largest resources for facial expression recognition but predominantly contains data from Western populations, potentially limiting its generalizability [7]. Similarly, the Japanese Female Facial Expression Database dataset (JAFFE) and the Extended Cohn-Kanade Database dataset (CK+) have been criticized for their limited demographic diversity, often overrepresenting specific demographic groups or lab-controlled settings [11]. Additionally, while performance metrics like accuracy and F1 scores are widely reported in FER research, these often fail to address fairness concerns. For instance, a survey by Li and Deng [12] provides comprehensive performance metrics for FER datasets, which we use to analyze trends, but notes the limited focus on fairness and demographic generalization. Recently, the release of AffectNet+ [14], an enhanced version of AffectNet, introduced soft emotion labels and automated demographic annotations for gender, age, and ethnicity, aimed at facilitating emotion recognition across diverse populations [14]. However, to the best of our knowledge, AffectNet+ is neither demographically

balanced nor widely accessible due to academic licensing restrictions, and the accuracy of its automated annotations has not been reported.

This project examines these challenges by addressing two key aspects of FER research: first, an analysis of trends in FER research datasets using a scraped dataset of academic papers to identify frequently cited datasets and their reported metrics from the survey by Li and Deng [12] and [8]; and second, an evaluation of demographic representation in accessible, frequently cited, and automatically annotated datasets to assess bias. While our work is limited in scope, it explores these aspects at a foundational level, inspired by prior research on FER dataset biases [7, 5, 11, 9].

Through this dual approach, we aim to investigate the relationship between dataset diversity and model performance, provide insights into the state of FER research, and highlight the importance of dataset diversity in creating fair and reliable systems.

Our contributions include a scraped dataset of FER research papers that serves as a reusable resource to analyze trends in this field, and a secondary image dataset containing prominent FER image datasets with automated demographic annotations, which serves as a tool to assess demographic bias.

2 Identifying Frequently Cited FER Datasets and Metrics

2.1 Dataset Identification

Our data collection process began with a predefined list of prominent FER datasets, sourced from Li & Deng's *Deep Facial Expression Recognition: A Survey* [12]. This study provided a comprehensive overview of frequently used FER datasets in high-impact research, which served as a foundation for our data scraping strategy. We aimed to gather scholarly articles referencing these well-cited datasets, focusing on their application in deep learning and FER studies.

2.2 Methodology

Our approach consisted of creating a dataset of scholarly articles through metadata extraction and enhancing it by retrieving and processing full-text content. A combination of APIs and web scraping techniques facilitated comprehensive data collection, enabling analysis.

2.2.1 Data Collection Process: Scholarly Search and Scraping

In our project, we utilized multiple data collection methods to gather relevant scholarly articles and full-text documents. The primary goal was to identify frequently mentioned FER datasets in high-impact research papers.

To collect high-impact scholarly articles, we employed an automated query mechanism leveraging Google Scholar through the Scholarly Python library [3]. This allowed us to extract metadata for papers citing popular FER datasets. Our search combined FER dataset names with specific topics like facial expression recognition, deep learning, and machine learning, as per the proposal's methodology. This iterative querying was essential for filtering papers based on citation counts, focusing on those with more than 100 citations. The collected data includes titles, authors, publication years, citation counts, detected datasets, detected topics, abstracts, digital object identifier (DOI), journal names, and URLs, which we saved in a CSV file for further analysis.

2.2.2 Web Scraping and Full-Text Retrieval

Given the limitations of Google Scholar for direct full-text retrieval, we expanded our data collection by implementing web scraping to download full-text documents for the datasets identified in the previous section, using the URLs extracted during metadata collection. Each source required specific handling due to differences in HTML structures and access protocols.

arXiv.org
Full-text PDFs were directly downloadable by extracting links with BeautifulSoup[10] after identifying the "View PDF" anchor on each publication page.

IEEE Xplore
To extract the full-text content from IEEE Xplore, we employed Selenium [15], a web automation tool. By utilizing Selenium to emulate a genuine user agent with Geckodriver [1], we were able to extract the DOI from the IEEE Xplore website. Subsequently, the extracted DOI was used to retrieve the corresponding PDF from a mirror website. Finally, the text content was extracted from the retrieved PDF using the PyPDF2 library.

ScienceDirect
To collect research articles from ScienceDirect [6], we developed an automated web scraper using Selenium for web automation. The scraper first logs in through institutional credentials to access restricted content. It then dynamically loads article pages, ensuring all elements are rendered. The scraper extracts and saves the HTML content for archiving purposes and converts the content to PDF format for flexibility across environments. Additionally, it extracts plain text using BeautifulSoup for content analysis. The scraper includes error handling mechanisms for timeouts, missing elements, and page structure changes to ensure reliable data collection.

2.2.3 API Usage

For more reliable data extraction, we prioritized the use of APIs over web scraping techniques, as APIs provide a structured method for requesting data and often return well-structured data in response. To retrieve metadata and citation information for scientific papers, we employed several APIs. General metadata, such as title, publication date, authors, URL, and abstract, were obtained using the Scholarly [3] Python library, which interfaces with the Google Scholar API, and the arXiv [1] API. Citation counts were retrieved using the CrossRef [4] API. However, for full-text retrieval, we resorted to web scraping

[1] `https://github.com/mozilla/geckodriver`

techniques, as the complete text of papers was hosted on various websites, each requiring specific handling and potentially necessitating the use of multiple APIs, if available.

2.2.4 Accuracy Scores Reported in Research

This subsection presents an accuracy score table compiled from the findings reported in prior studies, specifically [12] and [8]. The table consolidates performance metrics reported in these studies, providing a comparative overview of model performance on frequently referenced FER datasets 2 [IVAN APPENDIX ALSO]. By leveraging these reported results, we aim to highlight key trends and benchmarks in the field without conducting independent evaluations.

2.3 Results

In our analysis, the most frequently mentioned datasets in facial expression recognition (FER) research were identified based on the plot of top-detected datasets (Figure 1). These datasets include the Affective Faces Database (AffectNet), Toronto Face Database, Acted Facial Expressions in the Wild, MMI Facial Expression Database, Static Facial Expression in the Wild, Expression in-the-Wild, Japanese Female Facial Expression Database (JAFFE), Karolinska Directed Emotional Faces, Binghamton University 3D Facial Expression Database, and Extended Cohn-Kanade Database (CK+). Among these, we confirmed accessibility for CK+, JAFFE, and AffectNet. CK+ provides labeled expressions captured in a controlled environment, while JAFFE includes facial expressions from Japanese participants, primarily in a lab-controlled setting. AffectNet, the most frequently detected dataset in our analysis, features a large number of annotated facial images and has been widely utilized in the field.

Interestingly, shortly after the submission of our project description, an enhanced version of AffectNet, termed AffectNet+ [8], was released. This updated version introduces soft labels for emotions and demographic annotations for gender, age, and ethnicity, significantly expanding its utility for advanced FER tasks. However, AffectNet+ remains restricted to specific access conditions, requiring academic licensing or direct permission from its authors. Due to this limitation, we performed our own annotations on the original AffectNet dataset for demographic analysis.

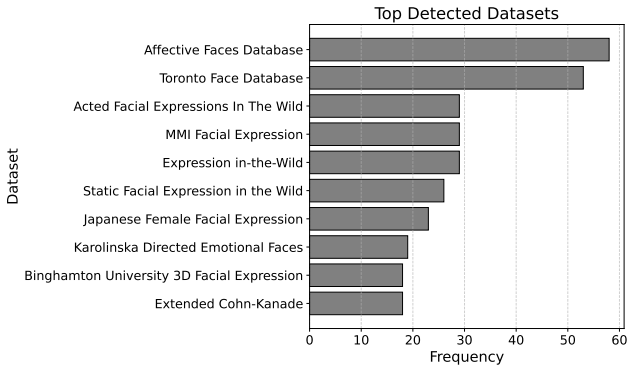


Figure 1: Most Frequently Detected Datasets in Facial Expression Recognition Research Papers

In the accuracy table reported in [**ivan˙table**], we identified that the FER2013 dataset was also accessible. Consequently, we included FER2013 in our project to facilitate additional demographic evaluations.

As shown in Figure 2, the accuracy ranges for frequently used FER datasets highlight notable differences in performance consistency, as reported by [12] and [8]. Datasets such as AffectNet+ and FER2013 exhibit broader accuracy ranges. In contrast, CK+ and JAFFE demonstrate narrower ranges.

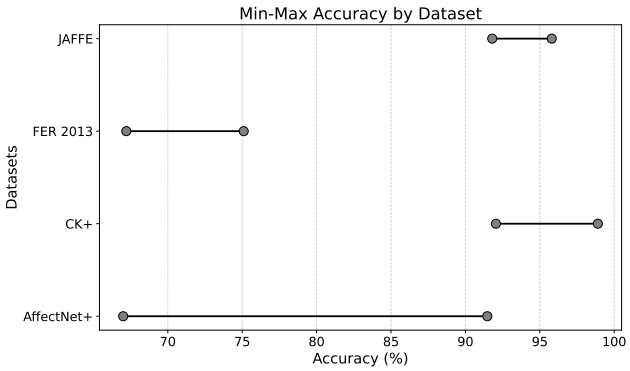


Figure 2: Range of Accuracy for Selected Datasets in Facial Expression Recognition

3 Evaluating Demographic Representation in FER Datasets

3.1 Objective

In this part of the project, our objective is to evaluate these datasets for demographic biases across ethnicities, age groups, and genders, which necessitates performing additional annotations due to the lack of labels beyond emotions.

3.2 Methodology

3.2.1 Dataset Selection and Annotation Workflow

Given the vast size of these datasets, we focused on subsets for automated annotation. From the AffectNet dataset, the test set containing 14,518 images was selected, while for FER2013, we used the test set comprising 7,178 images. The CK+ dataset (981 images) and the JAFFE dataset (213 images) were fully included due to their smaller sizes. In total, 22,890 images were selected for automated annotation.

It is important to note that the JAFFE dataset lacks explicit demographic variability in age, gender, and ethnicity, as it includes images solely of Japanese adult women, as indicated by its name ("Japanese Female Facial Expression"). The age range is typically from young to middle-aged. As a result, we are primarily interested in using JAFFE annotations to assess accuracy rather than demographic representation. Additionally, both FER2013 and CK+ are black-and-white datasets, which could pose challenges for annotation due to the absence of color information that often aids in identifying subtle facial features.

Despite this limitation, we proceed with these datasets to evaluate their utility in advancing FER research.

Automated annotation was conducted using the DeepFace [16] Python library, a comprehensive framework for facial attribute analysis. DeepFace leverages a hybrid architecture that combines insights from multiple state-of-the-art face recognition techniques to predict demographic attributes such as age, gender, and ethnicity. By integrating diverse model structures, DeepFace ensures adaptability and accuracy across various facial analysis tasks, offering functionalities such as face detection, alignment, and verification within a unified pipeline.

Age is initially provided as an integer but is converted into predefined age groups for consistency and ease of annotation. These groups include baby (0–1), child (2–9), adolescent (10–19), young adult (20–29), middle-aged adult (30–49), older adult (50–65), and elderly (65+). Gender is categorized into binary options, "Man" or "Woman," based on perceived appearance. Ethnicity is determined by selecting the dominant race from six categories: Asian, Indian, Black, White, Middle Eastern, and Latino/Hispanic. DeepFace was chosen for our automated annotation process due to its ease of use, wide adoption in the research community, and ability to predict multiple demographic attributes using a unified model. According to the original study, DeepFace achieved an accuracy of 73% for race and ethnicity prediction on the FairFace dataset, which consists of 86K training and 11K test instances with labels for seven different ethnicities. To evaluate the accuracy and reliability of this automated process, manual annotations were also performed.

In the **Manual Annotation** process, we followed DeepFace's predefined categories for gender and ethnicity to maintain consistency with the automated annotation process, while also adhering to the predefined age groups described above for demographic standardization. This approach ensures alignment with existing tools and simplifies the annotation workflow. While these categories could be expanded to include non-binary gender identities and more nuanced ethnic groups to better reflect real-world diversity, such an expansion was not feasible for automated annotation, so we adhered to the predefined categories. Due to resource constraints and the extensive manual effort required, only 42 images were annotated. This subset was carefully chosen to validate the reliability of automated annotations and serve as a benchmark for inter-annotator agreement. To ensure diversity, the images were selected to represent six ethnicity groups and seven age groups, with an equal distribution of male and female subjects (Appendix Figure A1). Despite efforts to represent all categories, the selection process required assigning each image to a category, making these annotations unsuitable for detailed statistical analysis. Three annotators participated in the manual annotation process, guided by a detailed annotation protocol (Appendix Figure .3). A custom-built annotation interface, implemented as a Python script, provided an interactive workflow with pop-up windows for annotations (Appendix Figure A2). Each annotator's results were saved as separate CSV files for subsequent analysis. Final labels were determined through a consensus-based approach, using majority voting to resolve discrepancies. Annotation consistency was evaluated using Fleiss' Kappa for categorical attributes (e.g., gender, ethnicity) and the Intraclass Correlation Coefficient (ICC) for ordinal data (e.g., age groups).

The final labels from the manual annotations are treated as the ground truth, while the automated system generates deterministic outputs (class labels). To evaluate the trustworthiness of automated annotations, we assessed accuracy and calculated Cohen's Kappa to measure agreement between the system's predictions and the ground truth, accounting for chance agreement. We also separately evaluated accuracy on the JAFFE dataset; however, in this case, the Kappa score would be zero because the dataset lacks demographic diversity, meaning there is no variability in key attributes (e.g., all subjects are Japanese women), making meaningful agreement calculations impossible.

3.3 Results

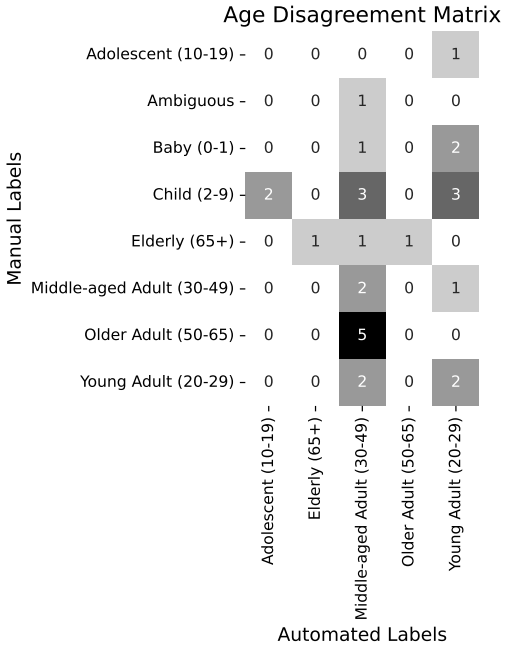
During the process of **manual annotation**, despite technical issues that resulted in the loss of annotations for four images from three annotators, 38 images were successfully annotated, ensuring the dataset's reliability for the majority of entries. For these annotations, inter-annotator agreement was evaluated using Fleiss' Kappa for two demographic categories: gender (0.78) and ethnicity (0.62). These values indicate substantial agreement for gender and moderate agreement for ethnicity. Age was evaluated using the Intraclass Correlation Coefficient (ICC), which yielded a value of 0.95, indicating excellent agreement among annotators. Out of a total of 126 labels (42 images across three categories), only two labels were found to be ambiguous. One case involved a baby where all three annotators provided different ethnicity labels, and another case involved a white woman for whom the annotators assigned different age groups as seen in Appendix A3. With a small sample size, ambiguity is more likely to arise from inherent challenges in the images rather than systematic biases among annotators. Overlapping characteristics in certain categories (e.g., age or ethnicity) introduce subjectivity in annotation. For the woman, subjectivity in age perception is a key factor, while for the baby, the lack of distinct ethnic features contributes to the disagreement. Since we do not have access to the true labels for these pictures, the best course of action is to accept the final manual annotation labels, including ambiguous ones, as the ground truth for this dataset. This approach represents the most reliable consensus available given the data.

The evaluation of **automated annotations**, summarized in Table 1, compares manual annotations (considered ground truth) with automated predictions across three attributes: Age, Gender, and Ethnicity.

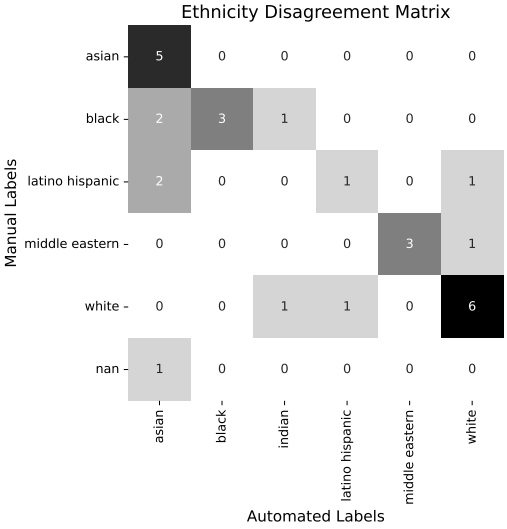
Table 1: Comparison of manual and automated annotations across attributes

Dataset	**Agreement (%)**	**Cohen's Kappa**
AffectNet (Age)	39.47	0.23
AffectNet (Gender)	73.68	0.40
AffectNet (Ethnicity)	73.68	0.68
JAFFE (Gender)	43.66	0.00
JAFFE (Ethnicity)	86.38	0.00

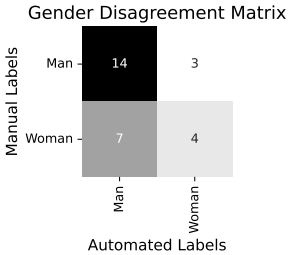
The results presented should be interpreted with caution due to the limited ground truth sample size, which may affect the reliability and generalizability of the findings. Age Agreement showed the lowest performance, with an agreement of 39.47% and a Cohen's Kappa of 0.23, indicating poor agreement. This suggests significant challenges in accurately predicting age categories, likely due to the inherent subjectivity and ambiguity in age perception from facial features. Looking at the mismatch matrix in Figure 3(a) The "Babies" and "Child" categories were not assigned at all during automated labeling, likely because DeepFace was not trained on these age groups. Additionally, the model struggles to differentiate between "Middle-aged Adult" and "Older Adult," possibly due to overlapping facial features and subtle age-related changes that are difficult for the model to discern. Gender Agreement achieved 73.68%,, with a Cohen's Kappa of 0.40, reflecting moderate agreement, while for JAFFE, Gender Agreement was 43.66%, with a Cohen's Kappa of 0.00, highlighting the system's limitations in accurately predicting gender, particularly for non-adult or ethnically homogeneous faces. Ethnicity Agreement for AffectNet was 73.68%,, with a Cohen's Kappa of 0.68, indicating substantial agreement, while JAFFE achieved an Ethnicity Agreement of 86.38% but a Cohen's Kappa of 0.00, likely reflecting the lack of diversity within the dataset. DeepFace demonstrates a tendency to misidentify women as men more frequently than vice versa, highlighting a potential bias in its gender classification. Additionally, the model shows the highest accuracy in correctly identifying individuals with White and Asian ethnicities, as evidenced by the low mismatch rates in these categories when comparing false and true labels. These findings underline the challenges automated systems face in predicting age and gender categories, especially for non-adult populations. However, the substantial agreement observed for ethnicity predictions, particularly in AffectNet, demonstrates relative strength in this category, making it a potential focus for further analysis of representation and bias across FER datasets.



(a) Age



(b) Ethnicity



(c) Gender

Figure 3: Disagreement Matrices for Automated Annotations

3.4 Bias Analysis and Ethical Considerations

We aimed to measure demographic attributes, including age, gender, and ethnicity, on FER datasets. The process of annotation posed significant challenges, ... also due to the presence of black-and-white images in a notable por-

tion of the datasets, which added complexity to automated labeling systems. To assess the accuracy of these automated systems, ground truth annotations were necessary. However, the limited number of annotations obtained in this study highlights the pressing need for large-scale annotation initiatives, potentially leveraging crowdsourcing. This approach could help increase the number of samples and annotators, improving dataset diversity and reliability.

Nevertheless, even with increased annotations, challenges such as annotator bias, subjective interpretations of demographic categories, and inconsistent labeling standards may persist. Despite these limitations, our results, as shown in the previous section, indicate that we can approximate the ethnicity distribution at a satisfactory level.

In Figure 4, it is evident that White individuals are overrepresented across all datasets, except for one dataset, which contains only Asian women.

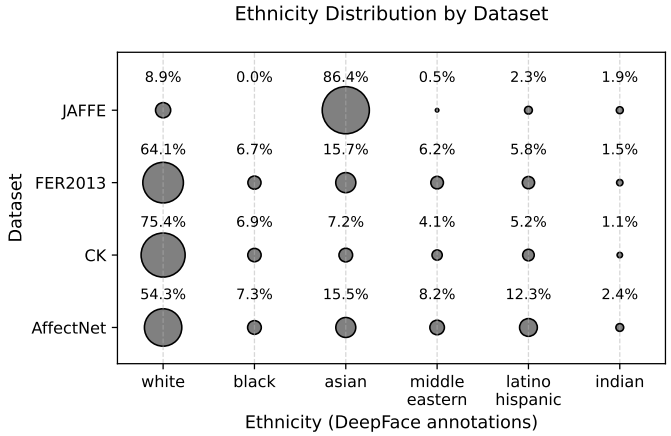


Figure 4: Ethnicity representation across FER datasets based on DeepFace predictions

These findings highlight a notable imbalance in the representation of demographic groups. [TODO: these results based on the DeepFace]

4 Discussion

We seek to analyze the impact of these biases on model performance and fairness, particularly when applied to populations that differ significantly from the datasets' dominant representation. This analysis is crucial to ensuring that FER models trained on these datasets can perform equitably across diverse demographics and cultural contexts.

In recent years, the FairFace[9] dataset has set a benchmark for addressing demographic imbalances in facial datasets. Designed with a specific focus on mitigating bias, FairFace ensures balanced representation across gender, race, and age categories. This approach highlights the importance of constructing datasets that promote fairness and inclusivity, addressing the limitations seen in many FER datasets that overrepresent specific demographics. Incorporating insights from datasets like FairFace can guide future efforts to create models that generalize effectively across diverse populations and minimize bias.

4.1 Table description draft

As a result of attaching the full text of the papers to the identifying information, we have effectively created a uniquely useful dataset for meta analysis of facial recognition papers. This table can be used for a number of NLP tasks such as searching through the text for different biases, sentimental meta analysis of the texts and comparing the methods used across the whole suite of papers. The data set also provides an easy look-up for which FER datasets are used and/or mentioned in the current research space for the subject. Another use could be to create an indexed knowledge base including directional links between the papers in order to maintain a strong overview of the subject matter.

4.2 Accuracy Score Tables

The accuracy scores and metrics in this study were drawn from authoritative surveys and studies, such as [12] and [8], rather than being extracted directly from the full text of our scraped data. This approach ensured credibility while providing a focused overview of performance benchmarks, aligning with the project's emphasis on demographic bias analysis and dataset evaluation. [ISN'T THIS SECTION REPETITIVE?]

4.3 Images

4.4 Impact of Demographic Diversity on Model Generalization

To evaluate the impact of demographic diversity on model performance, we combined insights from the above sources.

Our findings underscore the complex relationship between dataset diversity and model performance in FER tasks. Datasets with limited demographic variability, such as CK+ and JAFFE, tend to exhibit higher reported accuracy due to their controlled environments and homogeneity. However, these results may not generalize well to real-world applications where diversity in age, gender, and ethnicity is crucial. In contrast, more diverse datasets, such as AffectNet, show comparatively lower accuracies but provide a more realistic benchmark for FER models operating in varied demographic contexts. This dichotomy highlights the need for balancing accuracy with fairness and generalization in FER research, as also noted by Li and Deng [12].

The release of AffectNet+, with its demographic metadata and soft emotion labels, presents a significant advancement for the field. By leveraging soft labels, models can better capture the inherent ambiguity in human emotion expressions, potentially achieving higher accuracy while reflecting real-world complexity. However, the restricted access to AffectNet+ (requiring academic licensing or permission) limits its immediate impact and widespread adoption.

Our analysis reveals that datasets with limited demographic diversity, such as CK+ and JAFFE, tend to achieve higher accuracy but lack generalizability across diverse populations. In contrast, broader datasets like FER2013 face challenges in maintaining consistent performance due to their demographic diversity and real-world

complexity. The release of AffectNet+, with its demographic metadata and soft emotion labels, presents a significant advancement for the field. By leveraging soft labels, models can better capture the inherent ambiguity in human emotional expressions, potentially achieving higher accuracy while reflecting real-world complexity. However, the restricted access to AffectNet+ (requiring academic licensing or permission) limits its immediate impact and widespread adoption. This highlights the necessity for publicly accessible datasets that emphasize demographic diversity, akin to benchmarks like FairFace.

The small sample of manual annotations in our study reflects resource constraints but establishes a foundation for validating automated annotations. Future work could scale this process to bridge gaps in demographic fairness and improve generalizability across diverse populations.

5 Limitations and Future Work

Despite the above contributions, our work has several limitations.

In our study, we initiated our analysis with a predefined list of Facial Expression Recognition (FER) datasets, based on Li's survey [12]. However, this survey, published in 2018, did not include the FairFace dataset, which was introduced in 2021. Consequently, FairFace was excluded from our dataset selection. Since its release, FairFace has gained significant traction in the research community, as demonstrated by its growing number of citations and applications. This highlights the importance of conducting frequent and up-to-date surveys to capture emerging datasets and evolving trends in FER research.

Another limitation of this project lies in the refinement of the full-text data extracted from research papers. While basic cleaning was performed, the diverse formatting of PDFs introduced inconsistencies and artifacts in the final text. Issues such as the handling of tables, figures, and other structural elements remain areas for improvement to ensure seamless usage of the full texts.

The limited number of manual annotations—while providing valuable initial insights—restricts the scope of validating automated annotations. Scaling this process through larger datasets and leveraging crowdsourcing could improve demographic representation and enhance fairness. Additionally, reliance on automated annotations introduces potential inaccuracies due to biases in the underlying pre-trained models. Future work should focus on addressing these biases and validating findings through broader and more diverse annotation efforts.

6 Conclusion

In conclusion, this study highlights the critical role of demographic diversity in shaping the fairness and reliability of FER systems. By integrating insights from frequently cited FER datasets, demographic analyses, and recent advancements like AffectNet+, we emphasize the dual-edged nature of demographic labeling in FER research. While such datasets hold immense potential for improving fairness and performance, they must be employed responsibly to mitigate risks and ensure ethical outcomes. Addressing these challenges will be essential for advancing FER research and its applications in real-world settings.

Acknowledgments

The authors acknowledge the use of ChatGPT and Claude.ai for grammar enhancement in this paper. All technical content, analysis, and conclusions remain the author's original work.

References

[1] arXiv API. *arXiv API*. Computer software. 2024. URL: https://arxiv.org/help/api.

[2] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *FAT*. 2018. URL: https://api.semanticscholar.org/CorpusID:3298854.

[3] Steven A. Cholewiak et al. *SCHOLARLY: Simple access to Google Scholar authors and citation using Python*. Version 1.5.1. 2021. DOI: 10.5281/zenodo.5764801. URL: https://github.com/scholarly-python-package/scholarly.

[4] CrossRef API. *CrossRef API*. Computer software. 2024. URL: https://www.crossref.org/services/cited-by/.

[5] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. *Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition*. 2022. arXiv: 2205.10049 [cs.CV]. URL: https://arxiv.org/abs/2205.10049.

[6] Elsevier. *ScienceDirect*. Online database. 2024. URL: https://www.sciencedirect.com.

[7] Alex Fan, Xingshuo Xiao, and Peter Washington. *Addressing Racial Bias in Facial Emotion Recognition*. 2023. arXiv: 2308.04674 [cs.CV]. URL: https://arxiv.org/abs/2308.04674.

[8] Ali Pourramezan Fard et al. *AffectNet+: A Database for Enhancing Facial Expression Recognition with Soft-Labels*. 2024. arXiv: 2410.22506 [cs.CV]. URL: https://arxiv.org/abs/2410.22506.

[9] Kimmo Karkkainen and Jungseock Joo. "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 1548–1558.

[10] Leonard Richardson. *BeautifulSoup*. Python library for web scraping. 2024. URL: https://www.crummy.com/software/BeautifulSoup/.

[11] Shan Li and Weihong Deng. "A Deeper Look at Facial Expression Dataset Bias". In: *IEEE Transactions on Affective Computing* 13.2 (2022), pp. 881–893. DOI: 10.1109/TAFFC.2020.2973158.

[12] Shan Li and Weihong Deng. "Deep Facial Expression Recognition: A Survey". In: *IEEE Transactions on Affective Computing* 13.3 (July 2022), pp. 1195–1215. ISSN: 2371-9850. DOI: 10.1109/taffc.2020.2981446. URL: http://dx.doi.org/10.1109/TAFFC.2020.2981446.

[13] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Transactions on Affective Computing* 10.1 (Jan. 2019), pp. 18–31. ISSN: 2371-9850. DOI: 10.1109/taffc.2017.2740923. URL: http://dx.doi.org/10.1109/TAFFC.2017.2740923.

[14] Restack AI. *AI for Emotion Recognition: The Enhanced AffectNet+ Dataset*. https://www.restack.io/p/ai-for-emotion-recognition-answer-affectnet-dataset-cat-ai. Accessed: 2024-12-30. 2024.

[15] Selenium Project. *Selenium WebDriver*. Computer software. 2024. URL: https://www.selenium.dev.

[16] Sefik Serengil and Alper Özpınar. "A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules". In: *Bilişim Teknolojileri Dergisi* 17.2 (2024), pp. 95–107. DOI: 10.17671/gazibtd.1399077.

Appendix

Here is the content of the appendix.

.1 Accuracy Scores

Source [8]

Table 2: Accuracy Scores for Models Using AffectNet+

Model	**Emotion Class**	**Accuracy (%)**	**Methodology**
ResNet-50	Happy	85.86	Hard-FER (Easy)
ResNet-50	Sad	67.00	AU-Based Classifier
EfficientNet-B3	Fear	88.62	Binary Classifiers
XceptionNet	Disgust	91.47	Binary Classifiers
Ensemble	Contempt	78.51	Binary Classifiers

.2 Table

Datasets	Method	Network Type	Network Size	Pre-processing	Data Sele
Ouellet 14 [110]	CNN (AlexNet)	-	-	V	J
CK+	Li et al. 15 [86]	RBM	4	-	V
6 classes: 96.8					
CK+	Liu et al. 14 [13]	DBN CN	6	2m	-
CK+	Liu et al. 13 [137]	CNN, RBM CN	5	-	V
SVM 8 classes: 92.05 (87.67)					
CK+	Liu et al. 15 [138]	CNN, RBM CN	5	-	V
SVM 7 classes‡: 93.70					
CK+	Khorrami et al. 15 [139]	zero-bias CNN	4	7m	-
CK+	Ding et al. 17 [111]	CNN fine-tune	8	11m	IntraFace
CK+	Zeng et al. 18 [54]	DAE (DSAE)	3	-	AAM
CK+	Cai et al. 17 [140]	CNN loss layer	6	-	DRMF
CK+	Meng et al. 17 [61]	CNN MN	6	-	DRMF
CK+	Liu et al. 17 [77]	CNN loss layer	11	-	IntraFace
CK+	Yang et al. 18 [141]	GAN (cGAN)	-	-	MoT
CK+	Zhang et al. 18 [47]	CNN MN	-	-	-
JAFFE	Liu et al. 14 [13]	DBN CN	6	2m	-
JAFFE	Hamester et al. 15 [142]	CNN, CAE NE	3	-	-
MMI	Liu et al. 13 [137]	CNN, RBM CN	5	-	V
SVM 7 classes‡: 74.76 (71.73)					
MMI	Liu et al. 15 [138]	CNN, RBM CN	5	-	V
SVM 7 classes‡: 75.85					
MMI	Mollahosseini et al. 16 [14]	CNN (Inception)	11	7.3m	IntraFace
MMI	Liu et al. 17 [77]	CNN loss layer	11	-	IntraFace
MMI	Li et al. 17 [44]	CNN loss layer	8	5.8m	IntraFace
MMI	Yang et al. 18 [141]	GAN (cGAN)	-	-	MoT
TFD	Reed et al. 14 [143]	RBM MN	-	-	-
TFD	Devries et al. 14 [58]	CNN MN	4	12.0m	MoT
TFD	Khorrami et al. 15 [139]	zero-bias CNN	4	7m	-
TFD	Ding et al. 17 [111]	CNN fine-tune	8	11m	IntraFace
FER 2013	Tang 13 [130]	CNN loss layer	4	12.0m	-
FER 2013	Devries et al. 14 [58]	CNN MN	4	12.0m	MoT
FER 2013	Zhang et al. 15 [144]	CNN MN	6	21.3m	SDM
FER 2013	Guo et al. 16 [145]	CNN loss layer	10	2.6m	SDM
FER 2013	Kim et al. 16 [146]	CNN NE	5	2.4m	IntraFace
FER 2013	pramerdorfer et al. 16 [147]	CNN NE	10/16/33	1.8/1.2/5.3 (m)	-
SFEW 2.0	levi et al. 15 [78]	CNN NE	VGG-S	MoT	-
SFEW 2.0	Ng et al. 15 [63]	CNN fine-tune	AlexNet	IntraFace	-
SFEW 2.0	Li et al. 17 [44]	CNN loss layer	8	5.8m	IntraFace
SFEW 2.0	Ding et al. 17 [111]	CNN fine-tune	8	11m	IntraFace
SFEW 2.0	Liu et al. 17 [77]	CNN loss layer	11	-	IntraFace
SFEW 2.0	Cai et al. 17 [140]	CNN loss layer	6	-	DRMF
SFEW 2.0	Meng et al. 17 [61]	CNN MN	6	-	DRMF
SFEW 2.0	Kim et al. 15 [76]	CNN NE	5	-	multiple
SFEW 2.0	Yu et al. 15 [75]	CNN NE	8	6.2m	multiple

.3 Manual Annotation

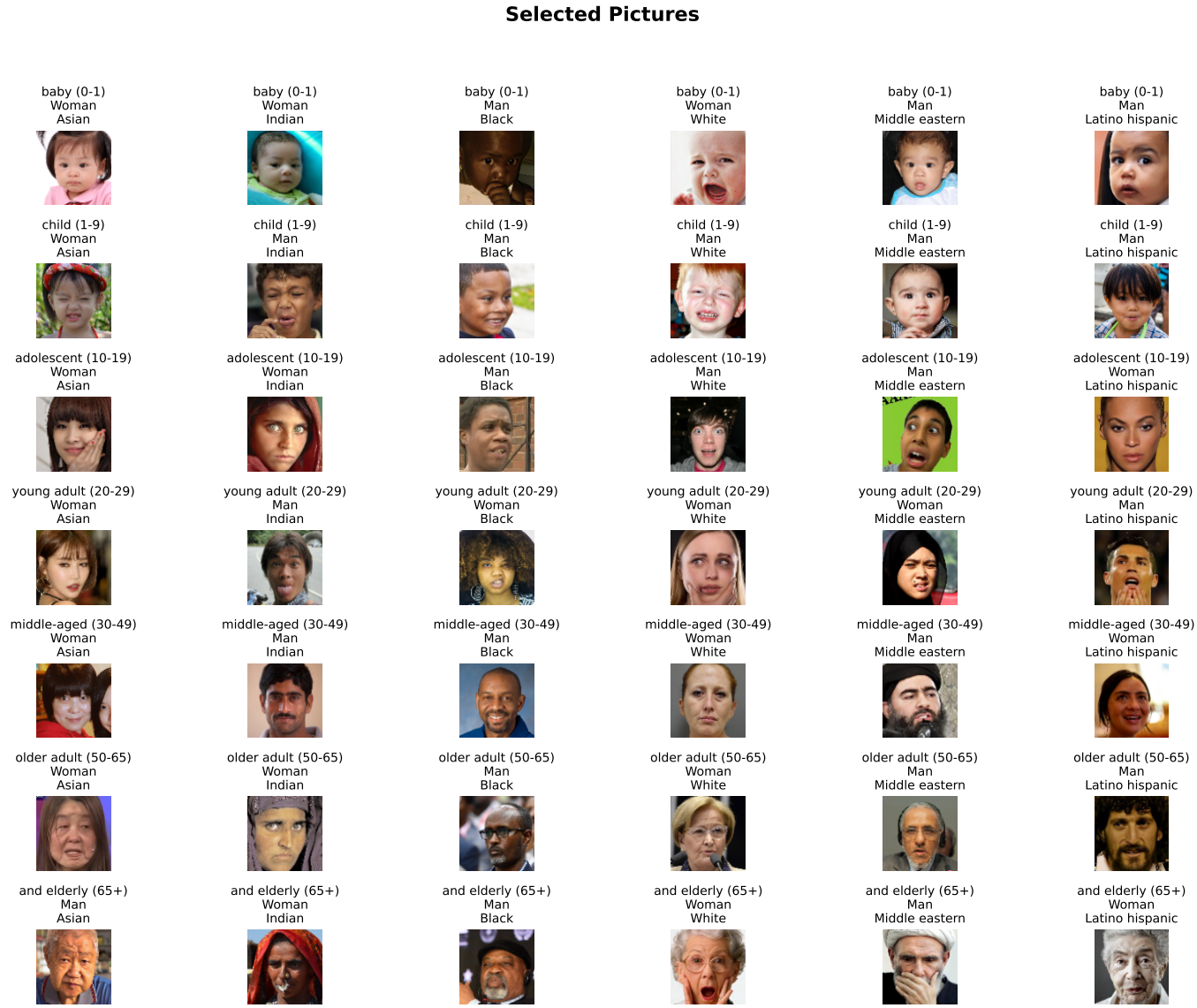


Figure A1: Selected Pictures for Manual Annotations

Annotation Guide

You will be required to annotate 42 images by providing answers for three categories: **age**, **gender**, and **ethnicity**.

Instructions

Step 1: Run the Code

Run the provided code. Two pop-up windows should appear.

Step 2: Follow These Steps

1. **Enter a Nickname**
   Begin by entering a nickname that cannot be linked to your identity. This ensures anonymity in the annotation process.

2. **Annotate Each Image**
   For each image, select one option for each category: **age**, **gender**, and **ethnicity**. Use the definitions below to guide your selection:

   - **Age Groups -** Select from the following options:

     - **Baby** (0–1 year)
     - **Child** (2–9 years)
     - **Adolescent** (10–19 years)
     - **Young Adult** (20–29 years)
     - **Middle-aged Adult** (30–49 years)
     - **Older Adult** (50–65 years)
     - **Elderly** (65+ years)

   - **Gender -** Choose **Man** or **Woman**. This category refers to perceived gender based on appearance.

   - **Ethnicity -** Choose from the following categories:

     - **Asian**
     - **Indian**
     - **Black**
     - **White**
     - **Middle Eastern**
     - **Latino/Hispanic**

   - *Base your judgments on visible traits. Keep in mind that these labels are for research purposes and may not fully capture individual identities.*

3. **Save Your Selections**
   After selecting an option for all three categories, click on **“Save.”**

4. **Proceed to the Next Image**
   Click **“Next”** to move to the next image.

5. **Repeat the Process**
   Continue annotating for all 42 images.

6. **Review Completed Annotations**
   After completing the process, review your entries to ensure no fields are left blank.

Important Notes

- **Accuracy and Judgment**
  Annotate each image as accurately as possible based on your best judgment. If uncertain, make your best guess. Your input is essential for identifying challenging cases and assessing annotation consistency.

- **Ground Truth and Consistency**
  Since we lack exact demographic information for the dataset, your annotations will help establish a "ground truth" based on consensus. This will aid in evaluating automated systems.

- **Ethical Considerations**
  Understand that categories like age, gender, and ethnicity involve subjective assessment and can be sensitive. This annotation task is designed to evaluate and improve automated systems, not to define individuals. Please approach the task with respect and awareness of the limitations of these categories.

- **Why Your Input Matters**
  Your annotations are critical for understanding consistency, identifying difficult cases, and improving automated systems. Your work contributes to building more accurate and fair models for the future.

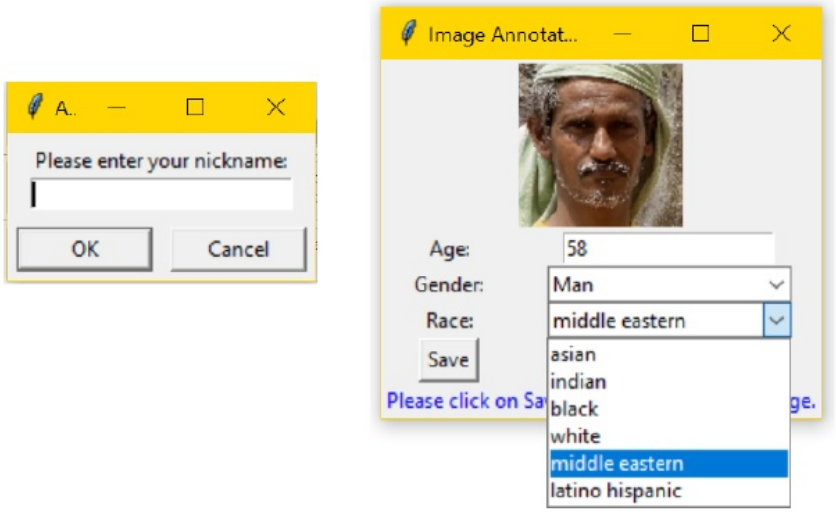


Figure A2: Interface for Manual Annotations

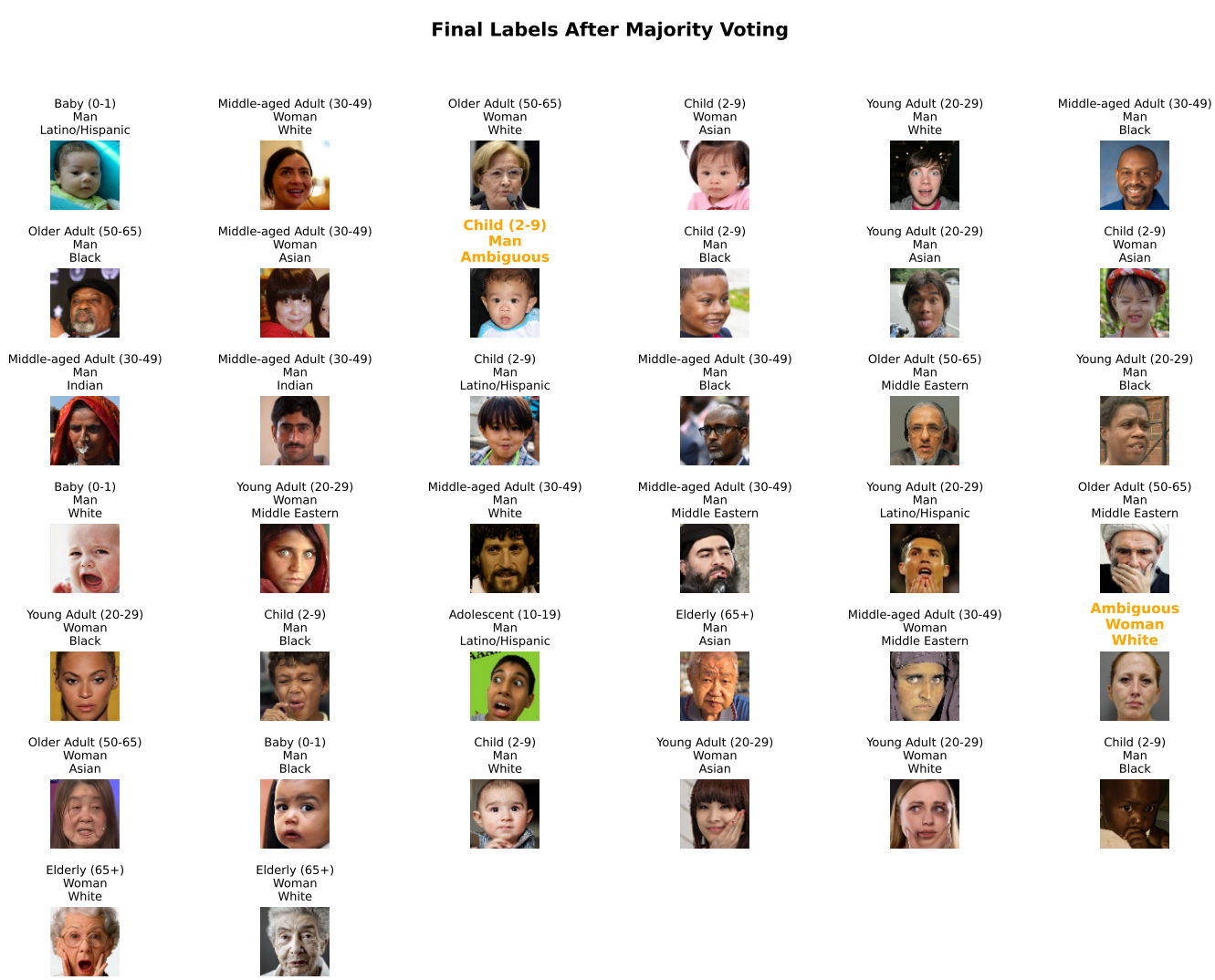


Figure A3: Manual Annotation Final Labels
only 2 ambiguous cases

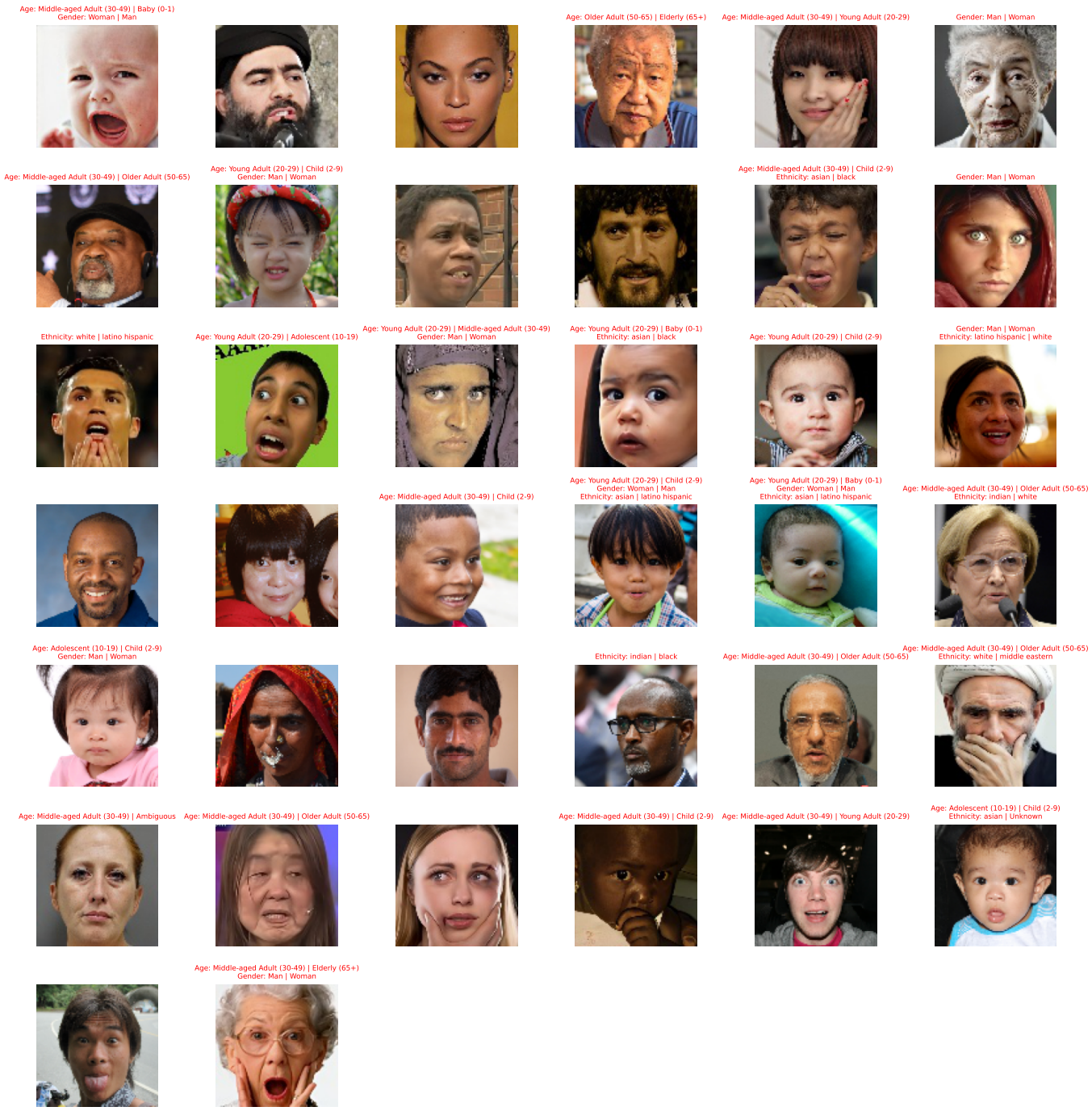


Figure A4: Automated vs. Manual Annotation:
Identifying Disagreements