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)

2025-01-02

**Abstract**

This study aims to achieve two objectives. First, we examine frequently cited Facial Expression Recognition (FER) datasets by creating a reusable dataset of 165 academic papers with metadata and full texts. Second, we evaluate the demographic diversity of four prominent FER datasets using automated annotation methods for age, gender, and ethnicity, and validate the results with manual annotations. Our findings reveal that datasets with limited demographic diversity tend to show higher accuracy ranges, while more diverse datasets exhibit broader accuracy ranges, starting from lower scores. This project underscores the importance of constructing equitable FER datasets and models by demonstrating the use of demographic annotations.

The necessary code and data to replicate this report are available on: Github

## 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 depend 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, 12].

While FER research continues to grow, many commonly used datasets face challenges related to demographic representation. The Affective Faces Database (AffectNet) [15] 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) [1] and the Extended Cohn-Kanade Database dataset (CK+) [14] have been criticized for their limited demographic diversity, often overrepresenting specific demographic groups or lab-controlled settings [12]. Additionally, although researchers widely report performance metrics like accuracy and F1 scores in FER research, they often overlook fairness concerns. For instance, a survey by Li and Deng [13] 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+ [16], 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 [16]. However, to the best of our knowledge, AffectNet+ is neither demographically balanced nor widely accessible due to academic licensing restrictions. Furthermore, the accu-

[1] Official site of the JAFFE dataset: `https://zenodo.org/records/3451524`

racy 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 [13] 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, 12, 10].

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.

To achieve these goals, we contribute two key resources that form the foundation of our analysis. 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* [13]. 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 improving 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 13 FER dataset names with 8 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. We selected 165 URLs for scraping from 261 unique papers, identifying 84 that IEEE Xplore, arXiv, and ScienceDirect published. We skipped retrieving 96 papers published across 26 different HTML structures due to the complexity and resource-intensive nature of the process.

**arXiv.org**

We directly downloaded full-text PDFs by extracting links with BeautifulSoup [11] after identifying the "View PDF" anchor on each publication page.

**IEEE Xplore**

To extract the full-text content from IEEE Xplore, we employed Selenium [17], a web automation tool. Using Selenium to emulate a genuine user agent with Geckodriver [2], we were able to extract the DOI from the IEEE Xplore website. Subsequently, we used the extracted DOI to retrieve the corresponding PDF from a mirror website. Finally, we extracted the text content 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. We obtained general metadata, such as title, publication date, authors, URL, and abstract, using the Scholarly Python library [3], which interfaces with the Google Scholar API [1], and the arXiv API [1]. We retrieved citation counts using the CrossRef API [4]. How-

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

ever, for full-text retrieval, we resorted to web scraping 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

We also used two accuracy score tables compiled from the findings reported in previous studies, specifically [13] and [8]. These tables consolidate performance metrics reported in these studies, providing a comparative overview of model performance on frequently referenced FER datasets (Appendix Table A.2 and A.3. 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, we identified the most frequently mentioned datasets in facial expression recognition (FER) research using the plot of top-detected datasets (Fig. 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. Our analysis frequently detected AffectNet as a dataset featuring a large number of annotated facial images, which researchers have widely used in the field.

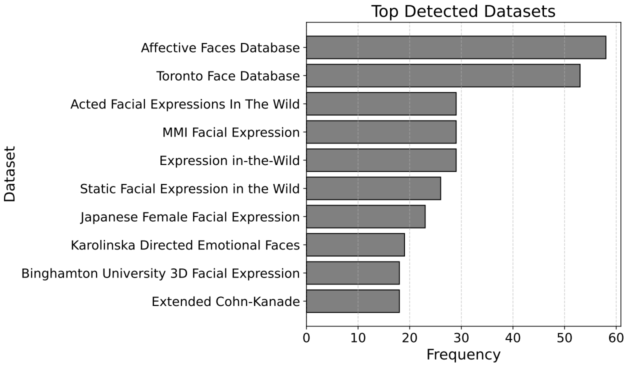


Figure 1: Most Frequently Detected Datasets in Facial Expression Recognition Research Papers

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, which contains the same images as AffectNet+ but with only one emotion label per image. So their demographic characteristics are the same.

In the accuracy table (Appendix Table A.3, we identified that the Facial Expression Recognition 2013 Dataset (FER2013) [9] dataset is also accessible. Consequently, we included FER2013 in our project to facilitate additional demographic evaluations, since it is also a frequently used dataset in facial expression recognition research.

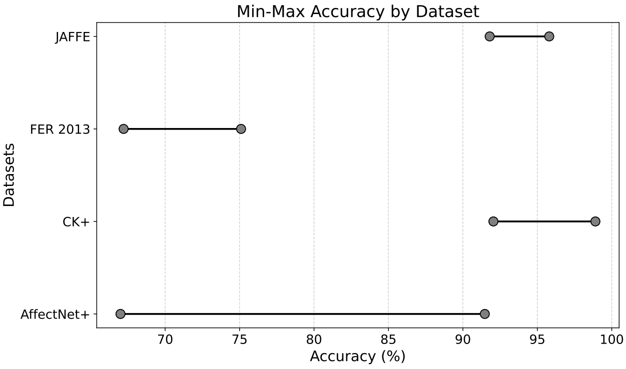


Figure 2: Range of Accuracy for Selected Datasets in Facial Expression Recognition

As shown in Fig. 2, the accuracy ranges for frequently used FER datasets highlight notable differences in performance consistency, as reported by [13] and [8]. Datasets such as AffectNet+ show a wide accuracy range, spanning low to high values, while FER2013 achieves predominantly low accuracy. In contrast, CK+ and JAFFE display narrower but consistently higher accuracy ranges.

# 3 Evaluating Demographic Representation in FER Datasets

## 3.1 Objective

In this part of the project, our objective is to evaluate the four datasets selected above for demographic biases across ethnicities, age groups, and genders, which requires 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, in two cases 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. We selected a total of 22,890 images for automated annotation.

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.

We conducted **Automated annotation** using the DeepFace Python library [18], a comprehensive framework for facial attribute analysis. DeepFace uses a combined design that includes insights from advanced 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. 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 [20], DeepFace achieved an accuracy of 68% 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, we also did manual annotations.

Age is initially provided as an integer but we have converted them 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+). DeepFace categorizes gender into binary options, "Man" or "Woman", and determines ethnicity by selecting the dominant one from six categories: Asian, Indian, Black, White, Middle Eastern, and Latino/Hispanic.

In the **Manual Annotation** process, we followed DeepFace's predefined categories for gender, ethnicity, and age groups for consistency with the automated annotation process. Although expanding these categories to include non-binary gender identities and more nuanced ethnic groups would better reflect real-world diversity, this was not feasible for automated annotation, so we retained the predefined categories for alignment. 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 Fig. 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 Fig. A.3). We implemented a custom-built annotation interface as a Python script, which provided an interactive workflow with pop-up windows for annotations (Appendix Fig. A2). Each annotator's results were saved as separate CSV files for subsequent analysis. We determined the final labels through a consensus-based approach, using majority voting to resolve discrepancies, and evaluated annotation consistency using Fleiss' Kappa for categorical attributes (e.g., gender, ethnicity) and the Intraclass Correlation Coefficient (ICC) for ordinal data (e.g., age groups).

We treat the final labels from the manual annotations 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.

We evaluated inter-annotator agreement 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 114 labels (38 images across three categories), we identified only two ambiguous labels. 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 (Appendix Fig. 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.

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.

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 |

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 Fig.3(a) The "Babies" and "Child" categories were not assigned at all during automated labeling, because DeepFace was not trained on these age groups [19]. 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.

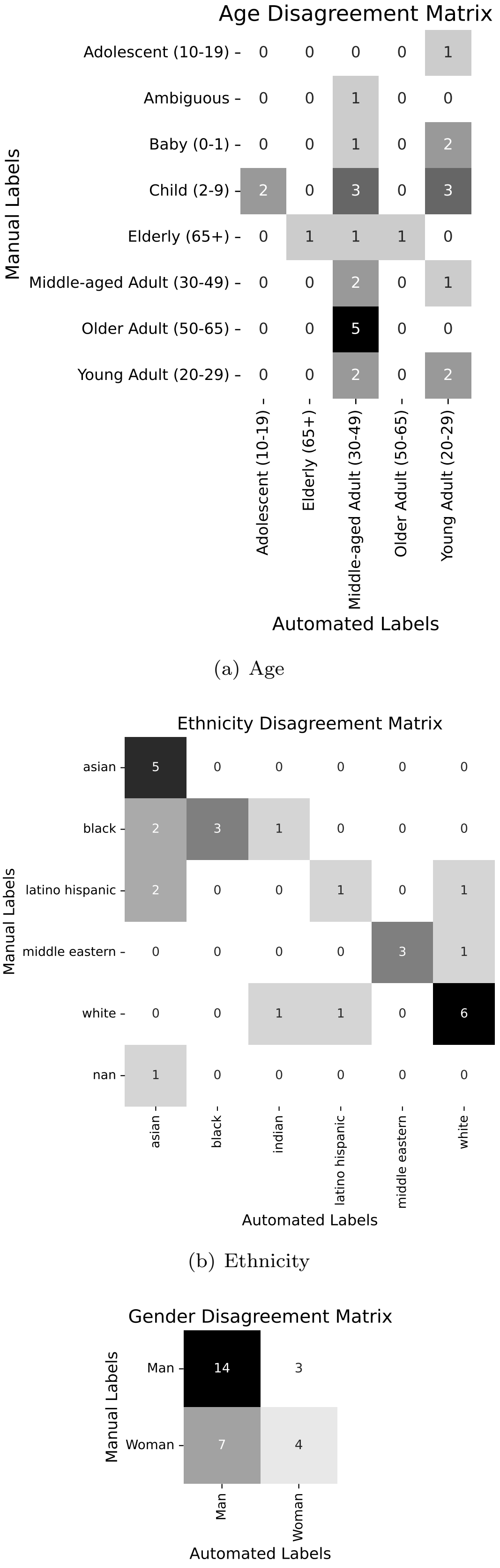
Gender Agreement achieved 73.68%,, with a Cohen's Kappa of 0.40, reflecting moderate agreement, while for JAFFE, Gender Agreement was 43.66%, with a Cohen's Kappa of 0.00, highlighting the system's limitations in accurately predicting gender, particularly for non-adult or ethnically homogeneous faces.

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.

Our ethnicity and gender agreement of 73% aligns with DeepFace's performance on FairFace (68%-72% [20]).

The substantial agreement observed for ethnicity predictions, particularly in AffectNet, demonstrates relative strength in this category, suggesting its potential for deeper analysis of representation and bias across FER datasets. Consequently, we evaluate only the ethnicity distribution within all four datasets, as shown in Fig. 4.

Fig. 4 reveals that White individuals are overrepresented across all datasets, with percentages ranging from 54.3% in AffectNet to 75.4% in CK+. We know JAFFE consists entirely of Asian women (100%), but DeepFace annotations identified this group as Asian women at a rate of (86.4%), showing that DeepFace also recognized this skewness. Other ethnic groups, such as Black, Middle Eastern, Latino/Hispanic, and Indian, are significantly under-represented, with proportions below 15.7% in all datasets. Among the datasets analyzed, AffectNet and FER2013 exhibit the highest diversity, although these datasets still reflect a high imbalance. These findings underscore a pronounced imbalance in demographic representation, highlighting the need for more diverse and inclusive FER datasets to ensure equitable model performance across populations.



(a) Age

(b) Ethnicity

(c) Gender

Figure 3: Disagreement Matrices for Automated Annotations

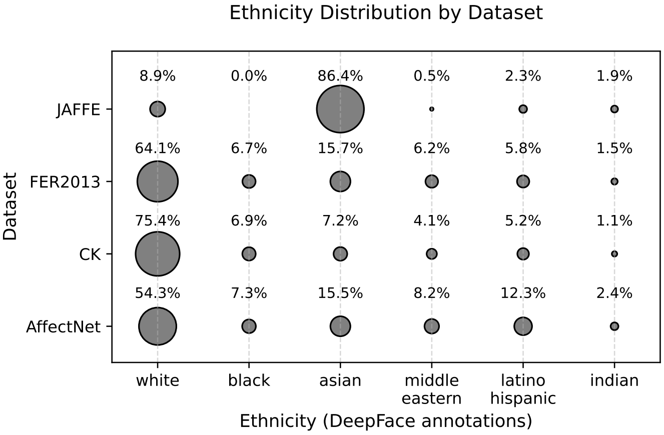


Figure 4: Ethnicity representation across FER datasets based on DeepFace predictions

# 4 Discussion

## 4.1 Scraped FER Research Paper Collection

As part of this study, we created a comprehensive dataset from 165 FER-related academic papers, including metadata such as titles, authors, publication years, citations, detected datasets, and extracted full texts. We used this resource to identify the most frequently cited datasets.

Beyond our application, the dataset can support bias identification, and meta-analysis to uncover patterns in FER methodologies and results. It also offers potential for creating an indexed knowledge base with directional links between papers, providing a cohesive view of FER literature and advancing understanding of the evolution, methodologies, and limitations of FER systems.

## 4.2 Images and Demographic Labeling

In the process of creating the secondary dataset with demographic annotations for images in FER datasets, our findings underscored the necessity of manual annotation to accurately evaluate the performance of automated methods. While we aimed primarily to focus biases in gender, age, and ethnicity annotation, the DeepFace results showed that only the ethnicity predictions from automated systems were reliable.

This validation is crucial, as automated labeling systems are often trained on similarly skewed datasets, keeping and spreading biases in their outputs. The annotation process further highlights the trade-offs between manual and automated methods. Manual annotations, while providing a more reliable result, are limited in scalability due to their resource-intensive nature. In contrast, automated methods, though efficient, inherit biases from pre-trained models, showing the need for constant checking and improvement to ensure fairness and reliability in FER systems.

## 4.3 Impact of Demographic Diversity on Model Generalization

Our analysis reveals that datasets with limited ethnic diversity, such as CK+ and JAFFE, tend to exhibit higher accuracy due to their controlled environments and demographic homogeneity. In contrast, slightly more diverse datasets, such as FER2013 and AffectNet, face challenges in maintaining consistent performance, reflecting the complexity introduced by their broader demographic representation and real-world variability. The introduction of soft emotion labels, as seen with the release of AffectNet+, offers a pathway to improve accuracy by capturing the nuanced nature of human emotions.

It is encouraging to see dataset creators increasingly prioritizing demographic insights, as demonstrated by AffectNet+ with its demographic metadata, or emphasizing demographic diversity, as with FairFace. These efforts mark a significant step toward addressing the biases inherent in FER datasets and improving the fairness and generalizability of FER systems.

# 5 Limitations and Future Work

Despite the above contributions, our work has several limitations.

In our study, we started our analysis with a set of Facial Expression Recognition (FER) datasets, based on Li's survey [13]. However, this survey, published in 2018, did not include the FairFace [10] dataset, introduced in 2021. Consequently, FairFace was excluded from our dataset selection. Since its release, FairFace has gained significant traction in the research community due to its balanced representation across gender, race, and age categories, setting a benchmark for addressing demographic imbalances. This highlights the importance of conducting frequent and up-to-date surveys to capture emerging datasets and guide efforts to create inclusive and equitable FER systems.

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.

Also, as future work, we could use the generated full-text dataset to identify the most frequently trained-on datasets, rather than just the most mentioned ones.

Furthermore, the limited number of manual annotations—while providing valuable initial insights—restricts the scope of evaluating automated annotations. The small sample in our study shows resource limits but builds a base for future validation work. Scaling this process through larger parts of datasets and leveraging crowdsourcing could improve usefully. However, even with increased annotations, challenges such as annotator bias, subjective interpretations of demographic categories, and inconsistent labelling standards may persist. Additionally, depending on automated annotations may cause errors 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

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 and conducting demographic analyses, we emphasize the necessity of demographic labeling in FER research, as well as the importance of validating the accuracy and trustworthiness of such labeling. While datasets with demographic annotations hold immense potential for improving fairness and promoting equity in systems that rely on FER, they must be utilized responsibly and with careful ethical considerations to avoid reinforcing existing biases.

# Acknowledgments

The authors acknowledge the use of ChatGPT and Claude.ai for grammar enhancement in this paper. ChatGPT also helped format and troubleshoot LaTeX. 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] Kaggle. *Challenges in Representation Learning: Facial Expression Recognition Challenge*. Accessed: 2024-11-15. 2013. URL: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.

[10] 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.

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

[12] 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.

[13] 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.

[14] Patrick Lucey et al. “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression”. In: *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE. 2010, pp. 94–101.

[15] 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.

[16] 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.

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

[18] 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.

[19] Sefik Ilkin Serengil. *Apparent Age and Gender Prediction in Keras*. Accessed: 2025-01-01. 2019. URL: https://sefiks.com/2019/02/13/apparent-age-and-gender-prediction-in-keras/.

[20] Sefik Ilkin Serengil. *Race and Ethnicity Prediction in Keras*. Accessed: 2024-11-15. 2019. URL: https://sefiks.com/2019/11/11/race-and-ethnicity-prediction-in-keras/.

# A Appendix

## A.1 AffectNet+ Accuracy Scores [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 |

## A.2 CK+, FER2013, JAFFE Accuracy Scores [13]

Table 3: Performance summary of representative methods for static-based deep facial expression recognition on the most widely evaluated datasets.
Network size = depth & number of parameters; Pre-processing = Face Detection & Data Augmentation & Face Normalization; NE = Network Ensemble; CN = Cascaded Network; MN = Multitask Network; LOSO = leave-one-subject-out

| Datasets | Network Type | Network Size | Pre- -processing | Data Group | Additional Classifier | Performance (%) |
|---|---|---|---|---|---|---|
| CK+ | CNN (AlexNet) | - / - | V&J | LOSO | SVM | 7 classes†: (94.4) |
| CK+ | RBM | 4 / - | V&J | LOSO | - | 6 classes: 96.8 |
| CK+ | DBN CN | 6 / 2m | ✓ | 8 folds | AdaBoost | 6 classes: 96.7 |
| CK+ | CNN, RBM CN | 5 / - | V&J | 10 folds | SVM | 8 classes: 92.05 (87.67) |
| CK+ | CNN, RBM CN | 5 / - | V&J | 10 folds | SVM | 7 classes‡: 93.70 |
| CK+ | zero-bias CNN | 4 / 7m | ✓ | 10 folds | - | 6 classes: 95.7;<br>8 classes: 95.1 |
| CK+ | CNN fine-tune | 8 / 11m | IntraFace | 10 folds | - | 6 classes: (98.6);<br>8 classes:(96.8) |
| CK+ | DAE (DSAE) | 3 / - | AAM | LOSO | - | 7 classes†: 95.79 (93.78)<br>8 classes: 89.84 (86.82) |
| CK+ | CNN loss layer | 6 / - | DRMF | 10 folds | - | 7 classes†: 94.39 (90.66) |
| CK+ | CNN MN | 6 / - | DRMF | 8 folds | - | 7 classes†: 95.37 (95.51) |
| CK+ | CNN loss layer | 11 / - | IntraFace | 8 folds | - | 7 classes†: 97.1 (96.1) |
| CK+ | GAN (cGAN) | - / - | MoT | 10 folds | - | 7 classes†: 97.30 (96.57) |
| CK+ | CNN MN | - / - | ✓ | 10 folds | - | 6 classes: 98.9 |
| JAFFE | DBN CN | 6 / 2m | ✓ | LOSO | AdaBoost | 7 classes‡: 91.8 |
| JAFFE | CNN, CAE | 3 / - | - | LOSO | - | 7 classes‡: (95.8) |
| FER 2013 | CNN loss layer | 4 / 12.0m | - | - | - | Test: 71.2 |
| FER 2013 | CNN MN | 4 / 12.0m | MoT | - | - | Validation+Test: 67.21 |
| FER 2013 | CNN MN | 6 / 21.3m | SDM | - | - | Test: 75.10 |
| FER 2013 | CNN loss layer | 10 / 2.6m | SDM | - | k-NN | Test: 71.33 |
| FER 2013 | CNN NE | 5 / 2.4m | IntraFace | - | - | Test: 73.73 |
| FER 2013 | CNN NE | 10/16/33 / 1.8/1.2/5.3m | - | - | - | Test:75.2 |

## A.3 Manual Annotation

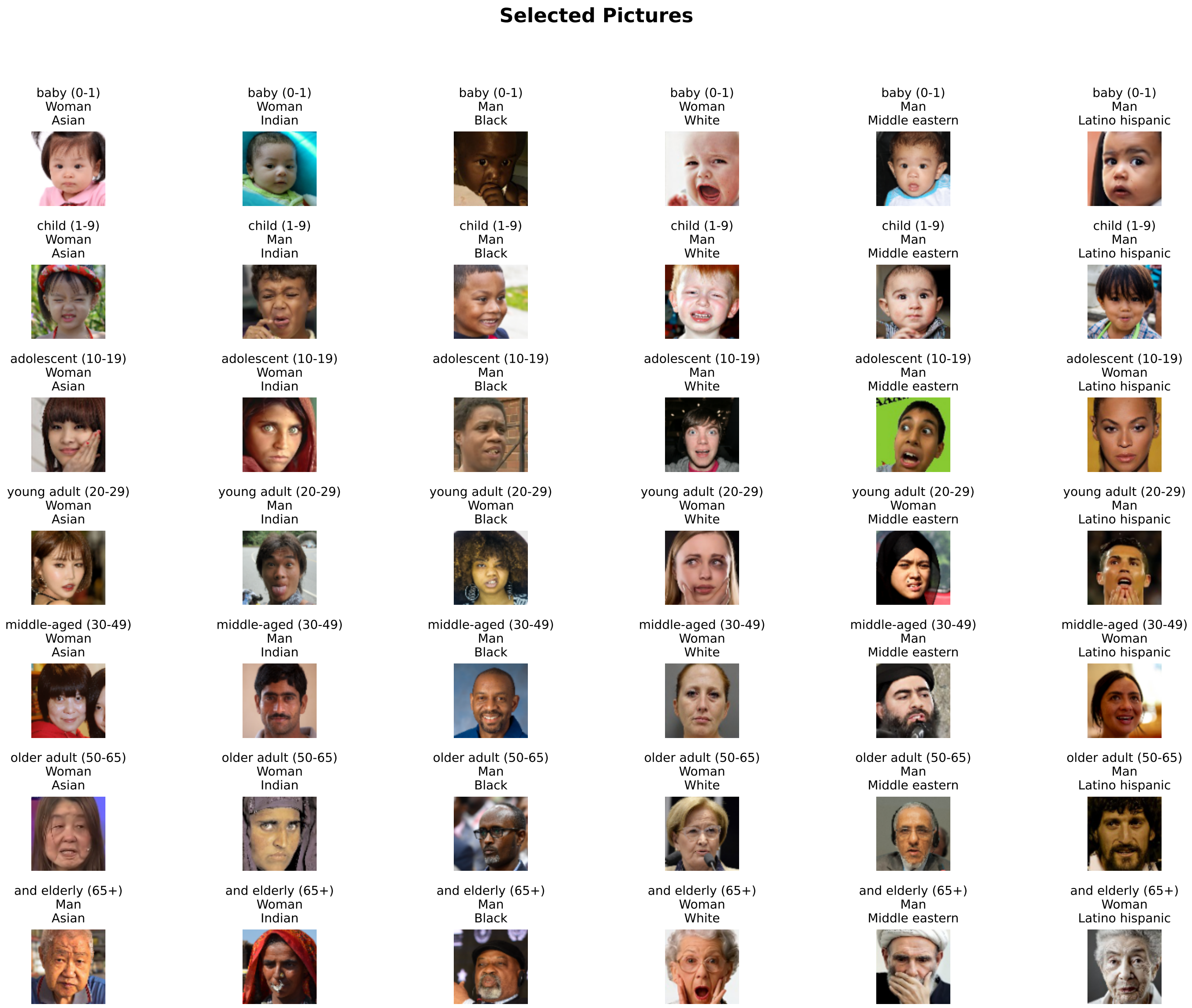
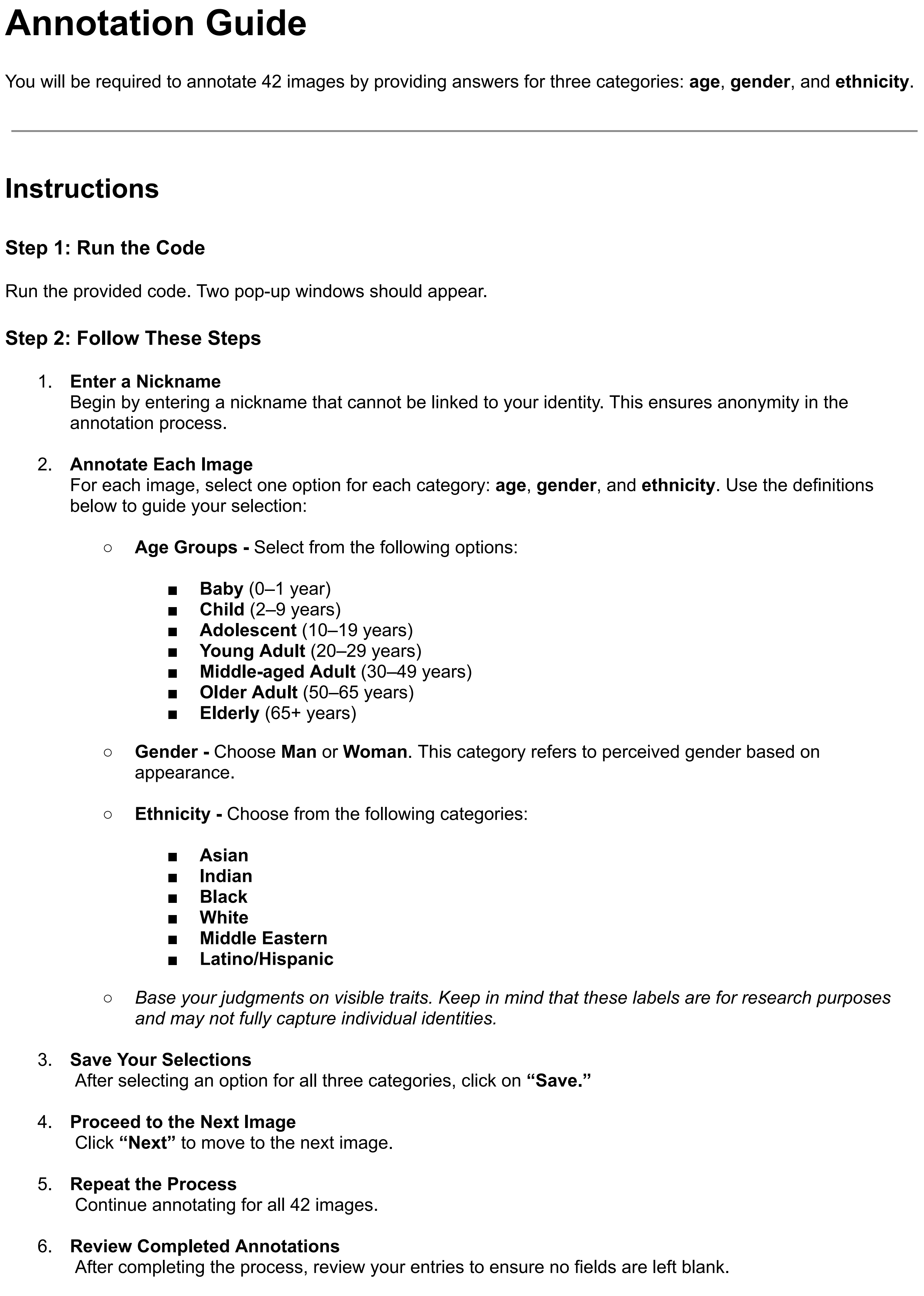

Selected Pictures
baby (0-1) Woman Asian
baby (0-1) Woman Indian
baby (0-1) Man Black
baby (0-1) Woman White
baby (0-1) Man Middle eastern
baby (0-1) Man Latino hispanic
child (1-9) Woman Asian
child (1-9) Man Indian
child (1-9) Man Black
child (1-9) Man White
child (1-9) Man Middle eastern
child (1-9) Man Latino hispanic
adolescent (10-19) Woman Asian
adolescent (10-19) Woman Indian
adolescent (10-19) Man Black
adolescent (10-19) Man White
adolescent (10-19) Man Middle eastern
adolescent (10-19) Woman Latino hispanic
young adult (20-29) Woman Asian
young adult (20-29) Man Indian
young adult (20-29) Woman Black
young adult (20-29) Woman White
young adult (20-29) Woman Middle eastern
young adult (20-29) Man Latino hispanic
middle-aged (30-49) Woman Asian
middle-aged (30-49) Man Indian
middle-aged (30-49) Man Black
middle-aged (30-49) Woman White
middle-aged (30-49) Man Middle eastern
middle-aged (30-49) Woman Latino hispanic
older adult (50-65) Woman Asian
older adult (50-65) Woman Indian
older adult (50-65) Man Black
older adult (50-65) Woman White
older adult (50-65) Man Middle eastern
older adult (50-65) Man Latino hispanic
and elderly (65+) Man Asian
and elderly (65+) Woman Indian
and elderly (65+) Man Black
and elderly (65+) Woman White
and elderly (65+) Man Middle eastern
and elderly (65+) Woman Latino hispanic

Figure A1: Selected Pictures for Manual Annotations

## Annotation Guide (Page 1)



# 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.

## Annotation Guide (Page 2)

# Important Notes

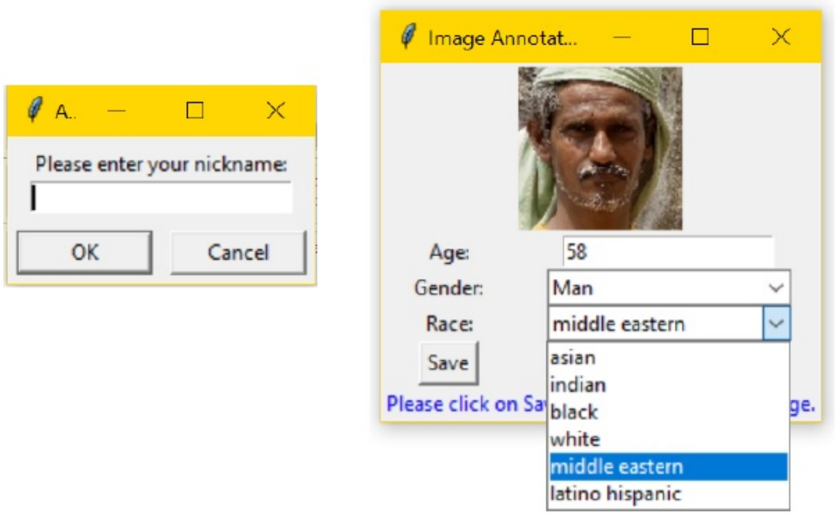
- **Accuracy and Judgment**
  Annotate each image as accurately as possible based on your best judgment. If uncertain, make your best guess. Your input is essential for identifying challenging cases and assessing annotation consistency.

- **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.

## A.4 Manual Annotation Interface



Please enter your nickname:

Age: 58

Gender: Man

Race: middle eastern

Figure A2: Interface for Manual Annotations

## A.5 Manual Annotation Final Labels

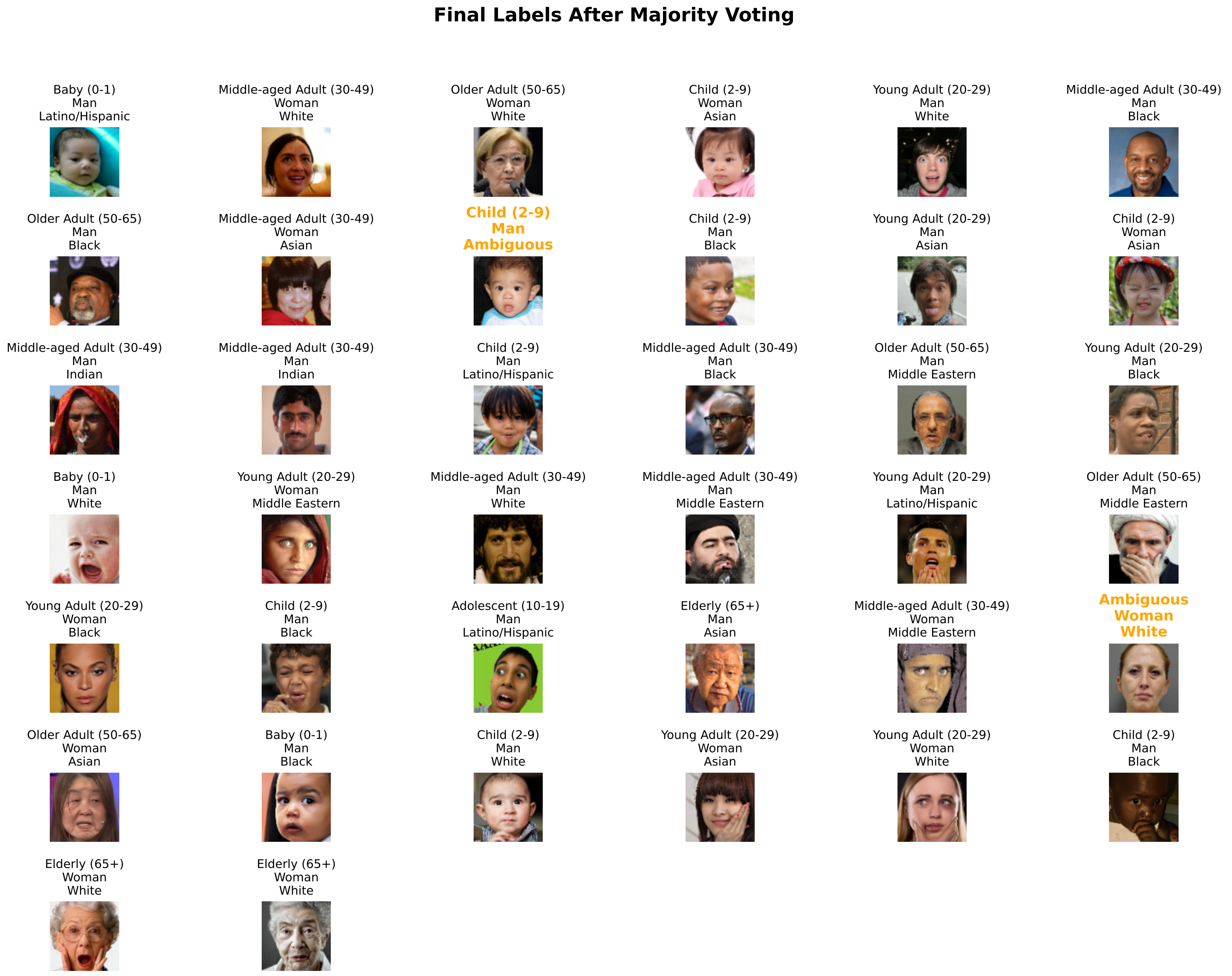

Final Labels After Majority Voting
Baby (0-1) Man Latino/Hispanic
Middle-aged Adult (30-49) Woman White
Older Adult (50-65) Woman White
Child (2-9) Woman Asian
Young Adult (20-29) Man White
Middle-aged Adult (30-49) Man Black
Older Adult (50-65) Man Black
Middle-aged Adult (30-49) Woman Asian
Child (2-9) Man Ambiguous
Child (2-9) Man Black
Young Adult (20-29) Man Asian
Child (2-9) Woman Asian
Middle-aged Adult (30-49) Man Indian
Middle-aged Adult (30-49) Man Indian
Child (2-9) Man Latino/Hispanic
Middle-aged Adult (30-49) Man Black
Older Adult (50-65) Man Middle Eastern
Young Adult (20-29) Man Black
Baby (0-1) Man White
Young Adult (20-29) Woman Middle Eastern
Middle-aged Adult (30-49) Man White
Middle-aged Adult (30-49) Man Middle Eastern
Young Adult (20-29) Man Latino/Hispanic
Older Adult (50-65) Man Middle Eastern
Young Adult (20-29) Woman Black
Child (2-9) Man Black
Adolescent (10-19) Man Latino/Hispanic
Elderly (65+) Man Asian
Middle-aged Adult (30-49) Woman Middle Eastern
Ambiguous Woman White
Older Adult (50-65) Woman Asian
Baby (0-1) Man Black
Child (2-9) Man White
Young Adult (20-29) Woman Asian
Young Adult (20-29) Woman White
Child (2-9) Man Black
Elderly (65+) Woman White
Elderly (65+) Woman White

Figure A3: Manual Annotation Final Labels
only 2 ambiguous cases

## A.6 Automated vs. Manual Annotation: Identifying Disagreements

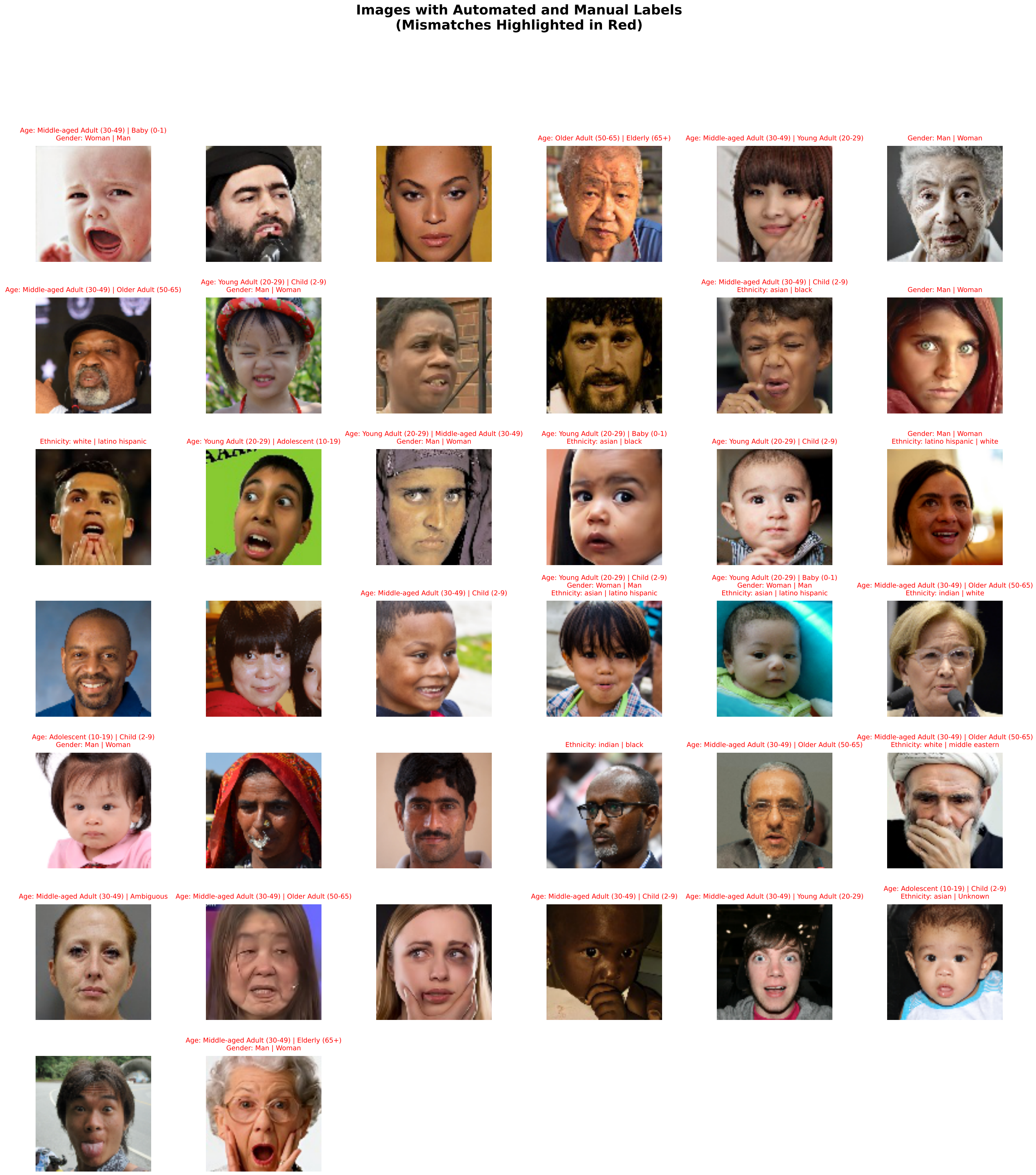

Images with Automated and Manual Labels
(Mismatches Highlighted in Red)
Age: Middle-aged Adult (30-49) | Baby (0-1)
Gender: Woman | Man
Age: Older Adult (50-65) | Elderly (65+)
Age: Middle-aged Adult (30-49) | Young Adult (20-29)
Gender: Man | Woman
Age: Middle-aged Adult (30-49) | Older Adult (50-65)
Age: Young Adult (20-29) | Child (2-9)
Gender: Man | Woman
Age: Middle-aged Adult (30-49) | Child (2-9)
Ethnicity: asian | black
Gender: Man | Woman
Ethnicity: white | latino hispanic
Age: Young Adult (20-29) | Adolescent (10-19)
Age: Young Adult (20-29) | Middle-aged Adult (30-49)
Gender: Man | Woman
Age: Young Adult (20-29) | Baby (0-1)
Ethnicity: asian | black
Age: Young Adult (20-29) | Child (2-9)
Gender: Man | Woman
Ethnicity: latino hispanic | white
Age: Middle-aged Adult (30-49) | Child (2-9)
Age: Young Adult (20-29) | Child (2-9)
Gender: Woman | Man
Ethnicity: asian | latino hispanic
Age: Young Adult (20-29) | Baby (0-1)
Gender: Woman | Man
Ethnicity: asian | latino hispanic
Age: Middle-aged Adult (30-49) | Older Adult (50-65)
Ethnicity: indian | white
Age: Adolescent (10-19) | Child (2-9)
Gender: Man | Woman
Ethnicity: indian | black
Age: Middle-aged Adult (30-49) | Older Adult (50-65)
Age: Middle-aged Adult (30-49) | Older Adult (50-65)
Ethnicity: white | middle eastern
Age: Middle-aged Adult (30-49) | Ambiguous
Age: Middle-aged Adult (30-49) | Older Adult (50-65)
Age: Middle-aged Adult (30-49) | Child (2-9)
Age: Middle-aged Adult (30-49) | Young Adult (20-29)
Age: Adolescent (10-19) | Child (2-9)
Ethnicity: asian | Unknown
Age: Middle-aged Adult (30-49) | Elderly (65+)
Gender: Man | Woman

Figure A4: Automated vs. Manual Annotation:
Identifying Disagreements