Facial Analysis Systems and Down Syndrome

Facial Analysis Systems and Down Syndrome

Marco Rondina(✉)[1], Fabiana Vinci[1,2], Antonio Vetrò[1], and Juan Carlos De Martin[1]

[1] Politecnico di Torino, Torino, Italy
{marco.rondina,antonio.vetro,demartin}@polito.it
[2] fabiana.vinci.fv@gmail.com

Abstract. The ethical, social and legal issues surrounding facial analysis technologies have been widely debated in recent years. Key critics have argued that these technologies can perpetuate bias and discrimination, particularly against marginalized groups. We contribute to this field by reporting on the limitations of facial analysis systems with the faces of people with Down syndrome: this particularly vulnerable group has received very little attention in the literature so far.
This study involved the creation of a specific dataset of face images. An experimental group with faces of people with Down syndrome, and a control group with faces of people who are not affected by the syndrome. Two commercial tools were tested on the dataset, along three tasks: gender recognition, age prediction and face labelling.
The results show an overall lower prediction accuracy in the experimental group, and other performance differences: i) high error rates in gender recognition in the category of males with Down syndrome; ii) adults with Down syndrome can be mislabelled as children; iii) social stereotypes are propagated in both the control and experimental groups, with labels related to aesthetics more often associated with females, and labels related to education level and ability more often associated with males.
These results, although limited in scope, shed new light on the biases that alter face classification when applied to faces of people with Down syndrome. They confirm the structural limitation of the technology, which is inherently dependent on the datasets used to train the models.

Keywords: Datasets · Face recognition · Face attribute estimation · Gender recognition · Age estimation · Image labelling · AI and disability · Down syndrome · AI bias.

1 Introduction and Motivation

In recent years, the ethical, social and legal implications of Facial Analysis Systems (FASs) have emerged in several parts of the world. Several municipalities and governments have banned the use of facial recognition technologies in public spaces, such as the city of San Francisco [17]. In Italy, a moratorium [14] suspended the use of these systems in public spaces and by private entities, except for activities related to criminal justice. The European Commission proposed

a restriction to 'real-time' remote biometric identification systems [4] while the European Parliament Research Services published an analysis on the regulation of facial recognition in the EU [5].

One of the main criticisms of FASs is their potential to perpetuate prejudices and discrimination. This is particularly the case for marginalized groups such as people of colour [20,2], or individuals with non-binary gender and transgender identities [11,19]. The amount of evidence and the implications of these issues are so significant that major companies have slowed development. Some have decided to remove gender prediction from their models [6] or to oppose the use of facial recognition technology for certain purposes [10].

This paper is part of a strand of research into the bias of FASs. It focuses on the limitations of FASs in relation to people with Down syndrome, an overlooked vulnerable group in this field. We built a dataset of images of people with and without Down syndrome (200 in each group, equally divided by binary gender) and used it to test and compare the classification of two commercial tools. The motivation behind this research was not to assess the resilience of technology and subsequently provide recommendations for enhancing classification. Our motivation was to provide new evidence on the structural limitations of FASs and their heavy dependence on training data. We did this from the perspective of a vulnerable social group that is often excluded from the design process and from the most popular discourses on AI bias and discrimination.

This paper is organized as follows: in Section 2 we position our work in relation to previous studies related to the topic of the research. In Section 3, we describe the design of our study, highlighting the research questions, the methodology used to create the test dataset and the details of the tested models. In Section 4 we present the results of our experiment for the three different tasks analysed, and the related discussion. Section 5 outlines the threats to validity and ethical concerns associated with the current study. Section 6 summarizes the reflections on the whole experiment. Section 7 explores possible future implementations. Finally, we have gathered the reference guidelines for the facial analysis services we selected, along with supplementary tables and figures, in the Supplementary Material[3].

2 Related Work

Several scholars have highlighted the ethical issues of FASs. Crawford in *Atlas of AI* [3] reconstructed the history and development of AI in relation to a variety of impacts (e.g., on the environment, work, health) and highlighted the epistemic issues of FASs and their controversial historical origins.

Other studies have focused on the failures of FASs and the associated negative impacts on society. In terms of gender and ethnicity, Buolamwini and Gebru [2] evaluated different commercial gender classification systems, and found that darker-skinned women were the most misclassified group. Similar results are reported by Klare et al. [9], who found that commercial and untrainable algorithms

[3] https://doi.org/10.5281/zenodo.8393211

performed worse for women, black people and young people. The issue of diversity and inclusion in FASs can arise from the lack of examples of sub-populations in the training dataset, but also from the definition of classes that incorporate specific values and beliefs, as in a binary formulation of gender [18]. The consequences of a non-inclusive operationalization of FASs can be seen in the case of transgender representation [8].

Recent work on the unknown behaviour of neural networks, has shown which are the key features used by commercial face classification services in order to classify gender. In fact, lip, eye, cheek structure and make-up are more discriminating than skin and hair length [12]. As the authors discuss, the fact that the make-up is so important in predicting female gender is a troubling stereotype.

Taking the next step and linking FASs to Down syndrome, several papers are based on detecting the disability in children in their first years of life. Agbolade et al. [1] presented a performance comparison of different machine learning methods on the task of Down syndrome detection. Paredes et al. [13] compared different machine learning and deep learning techniques to perform the emotion detection task on people with Down syndrome. Finally, Qin et al. [16] presented an identification method based on deep convolutional neural networks. The above studies aimed to identify the syndrome through the face or to better understand emotions through technology. What is not covered in the previous work is whether people with Down syndrome are discriminated against, in terms of lower FASs performance compared to non-affected people. This is the observable gap that we address in this paper.

3 Study Design

In this section we describe the research questions that drove the entire analysis (Section 3.1), how the test set was constructed (Section 3.2) and how the FASs were selected (Section 3.3).

3.1 Research Questions

RQ1: How do FASs, with images of Down-syndrome people, work with regard to predictions of a) gender and b) age?

Previous studies have demonstrated the failure of FASs to accurately predict age and gender for vulnerable individuals (Section 2). However, the same has not been investigated for people with disabilities. In this work, we are interested in understanding whether predicting gender and age for people with Down syndrome works in the same way as for people without the syndrome. The diverse consequences of inaccurate gender and age predictions are contingent on the decisions made regarding these predictions, such as hiring, incorrect associations for personalized contents.

RQ2: Do image recognition models assign labels differently to people with Down syndrome than to people without Down syndrome

The labels generated by the models and associated with an input image, can be used by companies for many purposes. They typically relate to the gender, objects and emotions depicted in the image. This study aims to scrutinize image labels and assess any potential variances between individuals with and without Down syndrome. These variations may result in unintended biases downstream, depending on the application context in which the FASs are employed.

3.2 Dataset

In the past, researchers have used datasets of faces of people with Down syndrome. Their experiments aimed to classify people with or without the syndrome [1,16]. However, they focus solely remained on children, who do not represent the entire population. It was therefore necessary to build a set of facial images of people with Down syndrome from scratch. This set serves as the experimental group (EG). Resource constraints prevented us from capturing images directly or reaching out to individuals for image contributions (with explicit consent) to assemble a sample. Thus, pre-existing images found on the internet proved the only available option. The images come from Google searches and from websites that offer free stock images, such as iStock and Pexels. In this way, it is impossible to know whether the individuals have given their explicit consent: for this reason, we err on the side of caution and do not redistribute them (see the discussion in Section 5). The resulting EG set consisted of 200 images.

The control group (CG), comprising of people unaffected by Down syndrome, consisted of 200 images selected from the UTKFace dataset[4] [21]. This dataset is a well-known large-scale face dataset with a long age range (from 0 to 116 years), constructed from images of famous people.

The dataset comprises 400 images in total, 100 for each of the following categories: EG male, EG female, CG male and CG female. Each image in the dataset was stored in two different ways, according to the reference rule of one of the tools used (see Section 3.3 that require cropped images, depicting solely the facial area, for detecting gender and age [5]). Thus, on the one hand, the *cropped* version of the images was used for gender and age recognition. On the other hand, the *not cropped* version of the images, was used for label detection.

All images within the dataset were paired with gender and age to evaluate model predictions. In accordance with the classifications used by the majority of FASs, gender is viewed as binary: we are aware of the limitations of such a representation. Age is the corresponding age of the person at the time the photo was taken. The information on *age* was the most difficult to find. For the EG, we were aware of the ages for 66 female and 64 male images. The images without age information were not used to predict age, but were used to predict gender and labels. It is also important to note that the life expectancy of people with Down syndrome is currently around 65 years [7]. Instead, the CG was created

[4] https://susanqq.github.io/UTKFace/

[5] this is the ClarifAI model, rule number 1, described in Appendix A.2 of the Supplementary Material

using a pre-existing dataset containing at least one image for each age from 3 to 85, so that all images had the age information.

One of the most important aspects for good model performance is the quality of the images. For this reason, we hand-picked all the images, resulting in the majority of the samples in the dataset conforming to the Pose, Illumination and Expression (PIE) rules [15].

3.3 Models

The services chosen for facial analysis were determined through a study of well-known commercial services that satisfy specific initial criteria:

- the images should not be retained for use for other purposes, or at a minimum be deleted with account suspension;
- it should be possible to predict gender and age;
- there should be a free amount of test operations.

The two services that met the previous criteria were: ClarifAI (CLAI) and Amazon Rekognition (AWSR) [6]. Some other services were considered, but not selected because they did not meet the above conditions. In particular: the services retain images indefinitely, as in the case of Face++ and Mega Matcher; they require a payment for the operations, Face++ and Cognitec's Face VACS; they do not provide gender recognition, Microsoft Face API, or they are deprecated, IBM Watson Visual Recognition.

Both ClarifAI and AWS Rekognition provide different models that are used during the analysis. The models and their details are shown and summarized in the Supplementary Material, Table 1, Appendix B. The selected services also provide some suggestions, called *Rules of reference*, for the correct use of models: they are summarized and reported in the Supplementary Material, Appendix A. The experiments were conducted using the SDKs for Python provided by AWSR and ClarifAI[7] [8]. The models generated a JSON file containing various types of information.

According to the research questions outlined in Section 3.1, this paper analyses three different tasks: gender recognition, age prediction and image recognition. As previously mentioned, gender is categorized as binary: male and female. Age prediction is performed differently by the two models. ClarifAI, assigns a probability to each of the possible age intervals for each image[9]: the predicted age interval is the one with the highest probability. Instead, AWSR predicts the

[6] regarding the first requirement, please refer to the policies on Appendix A in the Supplementary Material, in particular AWSR rule number 4, ClarifAI rule number 2

[7] https://docs.aws.amazon.com/rekognition/latest/dg/labels-detect-labels-image.html

[8] https://web.archive.org/web/20211130220210/https://docs.clarifai.com/api-guide/predict/images

[9] Table 2 of the Appendix B in the Supplementary Material.

Table 1: Gender recognition results.

		Experimental group				Control group			
Model	**Gender**	**Acc.**	**Prec.**	**Recall**	**F1-score**	**Acc.**	**Prec.**	**Recall**	**F1-score**
AWSR	*Female*	93%	87%	100%	93%	97%	94%	99%	96%
	Male		100%	85%	92%		99%	94%	96%
CLAI	*Female*	91%	87%	97%	92%	98%	97%	98%	98%
	Male		97%	85%	90%		98%	97%	97%

age by assigning to each image a specific range from a *'Low'* value to a *'High'* value. The outcome of this output format is that obtaining identical specific ranges for both models is unachievable. According to the rule of reference number 1 of the AWSR[10], the mathematical mean of the predicted range is taken as the final output value of age.

Image recognition models predict concepts, labels, themes and image properties. Analysis of this task provides insight into model training, specifically the labelling of images in the training datasets. ClarifAI provides a model (*general-image-recognition*), that outputs 20 different *concepts* for each image. Each *concept* has a corresponding probability value. The AWSR model (*detect labels*) assigns different labels to each image. We will refer to the *concepts* of ClarifAI as "labels" for the results of both models, ClarifAI and AWSR. A maximum of 20 labels are predicted for each image. For ClarifAI, the labels were categorized ad-hoc. Instead, the AWSR model automatically groups them by defining pre-existing categories (as specified in its documentation). All ClarifAI output labels were reviewed, analysed and categorizedaccording to their meaning. The final categories were *aesthetic*, *education*, *person description*. The labels of the first two categories are adjectives related to the people depicted in the photos. The *Person Description* category shares the same labels as the homonymous AWSR category. The similar categories of the AWSR model (*Clothing and Accessories*, *Beauty and Personal Care* and *Education*) contain mainly names of objects and descriptors of the images, which are not considered in the current analysis.

The peculiarity of the AWSR output labels is that they are mostly descriptive. The list of labels resemble a catalogue of objects recognized by the algorithm within the image. For example, considering one of its categories, *Apparel and Accessories*, some labels are: Jeans, T-shirt, Hat, Shoes etc. Instead, looking at the predicted labels from the ClarifAI model, it appears that the labels are predominantly adjectives. Adjectives can be much more ethically dangerous than nouns. Therefore, we focused our analysis on the *Aesthetics*, *Education*, *Person descriptors* categories of ClarifAI and on the category *Person descriptors* of the AWSR model.

[10] Appendix A.1 in the Supplementary Material

4 Results and Discussion

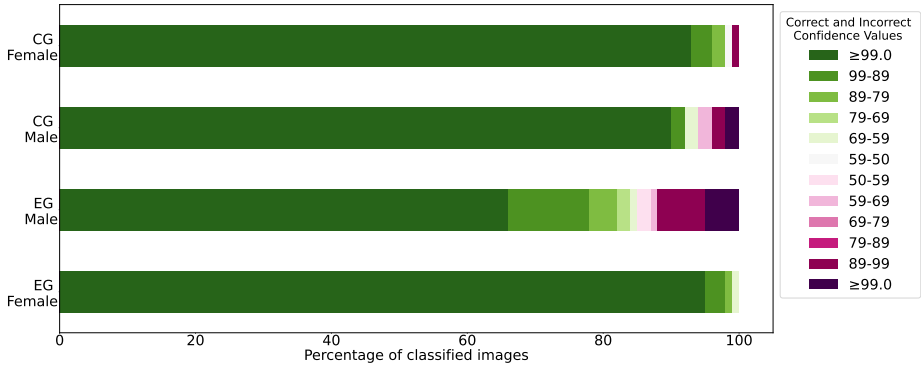
4.1 RQ1.a - Gender Recognition

Table 1 presents the values of *Accuracy*, *Recall*, *Precision* and *F1-score*, of both gender recognition models and groups. The accuracy scores of the Experimental Group (EG) were inferior to those of the Control Group (CG) for both models: the discrepancy between these scores was roughly 4% for Amazon Rekognition (AWSR) and 4% for ClarifAI (CLAI). Overall, the results of the EG were lower than those of the CG for both models. The *F1-scores* of the EG were lower than those of the CG. *Precision* for females and *Recall* for males showed lower values between the EG and the CG.

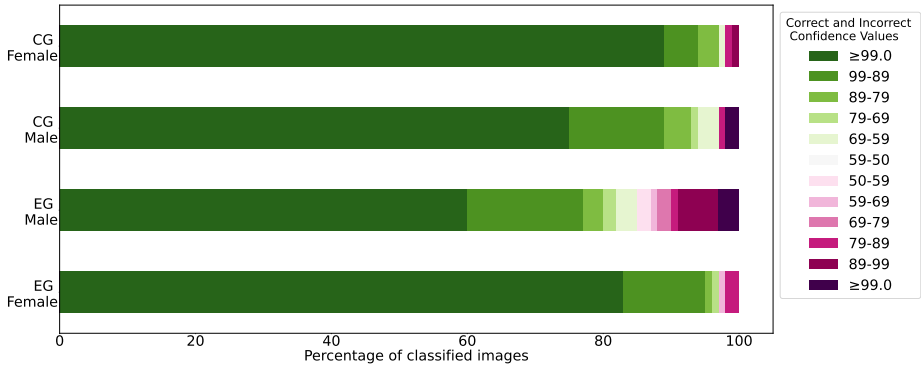
A closer examination of the misclassified images was carried out. On the one hand, all misclassified images of the EG male group represent children and adolescents. On the other hand, the misclassified images of EG females by the ClarifAI model represent old people. The rule of reference number 2 (Appendix A.1 in the Supplementary Material) regarding the AWSR model suggests that the confidence value assigned to each prediction of gender should be checked and taken into account. The threshold considered safe for sensible subjects is set at 99.00% by the rules of the model. Following the previous recommendation, a detailed examination of the confidence values is carried out on each group of the dataset, as shown in Figure 1.

Figures 1a and 1b illustrate the distribution of confidence values for each class of the dataset, including the values of correct and incorrect predictions: the colours green and purple represent the correctness and incorrectness of the model prediction in the study, respectively. The different shades of these colours represent the confidence levels and their own level of error. Both models performed poorly for the EG male category. The AWSR model correctly classified 66% of the images with a confidence level greater than or equal to 99.00%. The ClarifAI model correctly classified 60% of the images with a confidence level greater than or equal to 99.00%. The top performing categories were those belonging to females. In particular the category EG female, for the AWSR model, did not contain misclassification of gender. Looking at the incorrect predictions, we found that about a 5% of the predictions of the EG male classes had a confidence value greater than or equal to 99.00%. This indicates that the model made incorrect classifications even when it had high confidence in its predictions.

Finally, Table 2 shows the accuracy values considering only the prediction with a confidence value greater or equal to 99.00%. Within females, the discrepancy between EG and CG was reduced. In fact, for CLAI, the discrepancy between EG and CG is approximately 6%, whereas for AWSR, EG outperformed CG by 2%. The study reveals significant performance differences between the EG and the CG. Specifically, the accuracy difference was 24% in the AWSR case and 15% in the CLAI case.



(a) Confidence values computed by the AWSR model.



(b) Confidence values computed by the ClarifAI model.

Fig. 1: Confidence values for each category, computed by the AWSR and ClarifAI models, regarding the gender prediction task. Green bars refer to correct prediction, purple bars refer to incorrect predictions.

> We observed that theere was a discrepancy in gender prediction between the experimental group (EG) and the control group (CG). The study found a significant level of errors related to the EG male category. In particular, both models correctly predicted 85% of the images in the EG male class, in contrast to 97% and 94% of the CG male class for ClarifAI and AWSR respectively.

4.2 RQ1.b - Age Prediction

In the models analysed, age prediction is a classification problem and the result of the prediction corresponds to a range of ages instead of a precise value, as explained in Section 3.3. The accuracy values are presented in Table 3. The results show that the performances of both models are low: only half of the samples are predicted correctly. The truth tables were created using the ClarifAI

Table 2: Accuracy values for the gender prediction task, considering only the prediction with a confidence value greater than or equal to 99.00%.

	AWSR		**ClarifAI**	
	Experimental group	Control group	Experimental group	Control group
Female	95%	93%	83%	89%
Male	66%	90%	60%	75%

Table 3: Accuracy values for the age prediction task.

	AWSR		**ClarifAI**	
	Experimental group	Control group	Experimental group	Control group
Correct Age	52%	45%	48%	49%
Incorrect Age	48%	55%	52%	51%

(CLAI) ranges for both the true range and the predicted range. Each value of the truth tables 4a, 4b, 4c, 4d represent the number of images predicted in the corresponding range of ages. Looking at the CLAI model (Tables 4b and 4a) we can observe some differences between EG and CG. The EG performed worst in the ranges: 20-29, 30-39, 40-49, while the CG performed worst in the ranges between: 40-49, 50-59, 60-69, $\geq$ 70. The Amazon Rekognition (AWSR) predictions (Tables 4d and 4c) were slightly more accurate.

Most of the errors in the EG fell within the ranges of 30-39 and 40-49, while the CG had errors in the ranges of 30-39, 60-69 and$\geq$ 70. The comparison must consider that individuals with Down Syndrome have an average life expectancy of approximately 65 years. Additionally, it is noteworthy that the images in the EG dataset portray individuals with a maximum age range of 50-59.

Focusing on CLAI's predictions for the EG, Table 4a shows that the predicted ranges 3-9 and 10-19 are the two intervals with higher variance in the dataset. Thus, some images with a true age ranges of 20-29 and 30-39 were labelled with the age range 3-9 or 10-19. Similar outcomes can be observed for the age range of 40-49 with predictions of 10-19. Differently, for the CG, Table 4b shows that the age ranges of 20-29 and 30-39 have higher variance. Furthermore, certain images with a true age range of 50-59 or 60-69 were labelled incorrectly as being from the age ranges of 20-29 and 30-39. For the CG, the predicted age ranges are lower than the real ones, but they never coincide with the age ranges of children. The lower age predictions could be attributed to the dataset comprising images of well-known individuals who have undergone facial surgery and makeup to appear younger. Conversely, images of adult people with Down syndrome are classified within the child age ranges of 3-9 and 10-19.

The AWSR model is more stable in the ranges of its predictions. The ranges displaying greater variance for the EG are 3-9 and 10-19, while those for the CG exhibit almost uniform variance across all ranges. Additionally, certain errors in the AWSR model closely resemble those in the CLAI model for the EG. Some

Table 4: Truth tables regarding age prediction.

(a) ClarifAI Experimental group.

Experimental group									
	PREDICTED RANGE								
TRUE RANGE	**0-2**	**3-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**>70**
0 - 2	2								
3 - 9	3	9							
10 - 19		4	15	9					
20 - 29		3	14	24	10				
30 - 39		2	2	9	6	3			
40 - 49			1	2	2	6	2		
50 - 59						1	0	1	
60 - 69									
>70									

(b) ClarifAI Control group.

Control group									
	PREDICTED RANGE								
TRUE RANGE	**0-2**	**3-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**>70**
0 - 2	0								
3 - 9		11	4						
10 - 19		5	13	7					
20 - 29			2	20	3				
30 - 39				9	15				
40 - 49				4	9	11	3		
50 - 59				3	6	9	10	3	
60 - 69					3	4	12	9	
>70				1			7	11	6

(c) AWSR experimental group.

Experimental group									
	PREDICTED RANGE								
TRUE RANGE	**0-2**	**3-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**>70**
0 - 2	1	1							
3 - 9	1	10	1						
10 - 19		3	14	9					
20 - 29			15	22	12	2			
30 - 39		1	2	10	9				
40 - 49				2	5	4	2		
50 - 59						1	1		
60 - 69									
>70									

(d) AWSR control group.

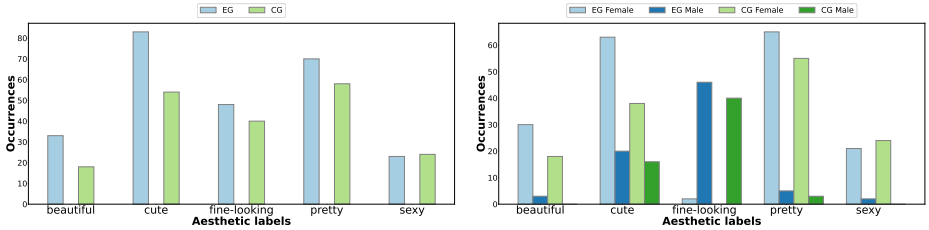
Control group									
	PREDICTED RANGE								
TRUE RANGE	**0-2**	**3-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**>70**
0 - 2	0								
3 - 9	1	9	5						
10 - 19		3	12	10					
20 - 29			3	21	1				
30 - 39			1	10	7	3			
40 - 49				2	6	17	2		
50 - 59					2	12	17		
60 - 69						4	15	6	1
>70							6	15	3

images with a true age range of 30-39 are assigned an age range of either 3-9 or 10-19.

We observed that there are some differences in age estimation between the experimental group (EG) and the control group (CG). The results lead to the conclusion that both models assign the age range of children to adults belonging to the experimental group (EG).

4.3 RQ2. Image Labelling

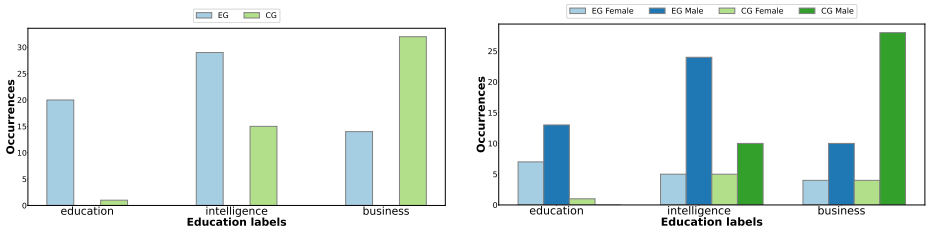
Aesthetics and Education Figure 2a illustrates that the experimental group (EG) had a greater number of instances in comparison to the control group (CG) for all concepts except for the label *Sexy*, although the difference is minimal. Looking at the gender distinction, Figure 2b, it is noticeable that for both the EG and the CG, women were more likely to be associated with the aesthetic labels than men. Females and males were assigned 316 and 135 labels, respectively. Furthermore, the label *Sexy* is exclusively attributed to images that were identified as female in the gender recognition task. The description of this label is linked to the ability to attract sexual desire or interest, which is associated exclusively with the female gender. With respect to the labels under the category of *Education*, Figure 3b, it is notable that the majority of the labels were linked with images depicting males rather than females. The overall situation reflects



(a) Comparison between experimental group and control group.

(b) Gendered comparison between experimental group and control group.

Fig. 2: Comparison regarding *Aesthetic* labels assigned by the ClarifAI model



(a) Comparison between experimental group and control group regarding.

(b) Gendered comparison between experimental group and control group.

Fig. 3: Comparison regarding *Education* labels assigned by the ClarifAI model.

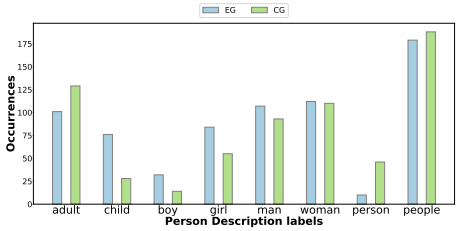
the conventional gender stereotypes, with women receiving aesthetic labels and men receiving educational labels.

Person Descriptors The name *Person description* is taken from one of the predefined categories of the AWSR model. All the labels are common to both models, so a comparison can be made as shown in Figure 4a and Figure 4b.

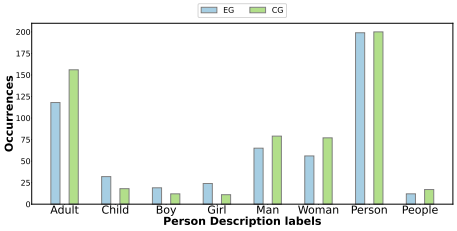
Both models assigned the label *Child* more often to the EG than to the CG, although the number of images representing children is quite balanced between the two groups. The same consideration can be made for the label *Adult*, where the situation is reversed: this result is consistent with the observations from the age prediction task, i.e. the models were more likely to consider a person in the EG as a child rather than an adult.

Each image represents only one person, and according to the definition given by the CLAI model for the label *Person* - one human being - and the label *People* - (plural) any group of people (men or women or children) together - the label *Person* should have been the correct one for each image in the dataset. CLAI correctly labelled only 56 images out of a total of 400 images, whereas AWSR correctly labelled 399 images out of a total of 400 images.

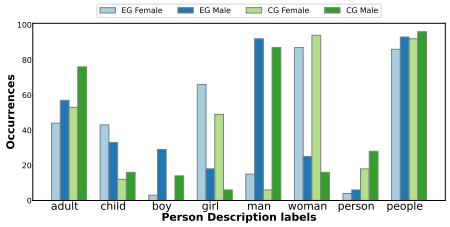
The other labels are gendered in the sense that they refer strictly to one gender rather than the other. For this reason, it may be useful to look at the



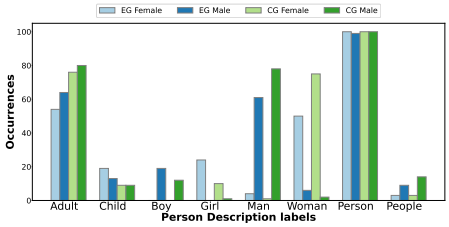
(a) Comparison between experimental group and control group regarding *Person Description* labels assigned by the ClarifAI model.



(b) Comparison between experimental group and control group regarding *Person Description* labels assigned by the AWSR model.



(c) Gendered comparison between EG and CG regarding *Person Description* labels assigned by the ClarifAI model.



(d) Gendered comparison between EG and CG regarding *Person Description* labels assigned by the AWSR model.

Fig. 4: Comparison regarding *Person Description* labels assigned by the ClarifAI and AWSR models.

Figure 4c and 4d where a comparison is made of the labels assigned between all four different classes of the dataset: EG male, EG female, CG male, CG female. Gendered labels were assigned to both genders, resulting in some female images being labelled as *Man* and vice versa. Some images had both labels *Man* and *Woman* or *Boy* and *Girl*. Both of these considerations reflect a general confusion in the assignment of these types of labels.

The concept of gender as binary is a product of choices made when defining labels, and is influenced by particular values and beliefs. Some labels in the *Person Description* category, such as *Male*, *Female*, *Man*, *Woman*, etc., carry the potential to categorize individuals based purely on their physical traits. Incorrect gender labelling can have a number of negative consequences for people, especially for groups that are systematically discriminated against. It is therefore questionable whether the use of gender labels is appropriate.

The results obtained by the image labelling models do not show significant differences between the EG and the CG. However, they highlight considerable gender disparities. In particular, the labels in the *Aesthetics* category are more likely to be associated with female classes than with male classes. Conversely, the labels of the *Education* category are more likely to be associated with male classes than with female classes.

5 Threats to Validity and Ethical Concerns

Most of the limitations and ethical concerns of the study relate to the construction of the dataset and the choice of models. Starting from the limitations of the dataset, some images are of poor quality. The process of resizing the images to obtain the cropped versions unavoidably diminished the quality of roughly 15% of the images associated with the control group (CG). Although the influence of the decreased quality on the classification outcomes may not be disregarded, it affected only a small number of images in the collection.

As described in Section 3.2, some images of the experimental group (EG) did not include information about the year in which they were taken or the age of the person depicted. The resulting limitation is an unbalanced comparison between the two groups, EG and CG, in terms of the different number of samples for each age range.

Since there is no information and no details on which datasets were used for the learning phase of the models, we cannot exclude the possibility that the same images, or some of them, were used in the training process. Building a test set with homemade images requires relevant economic and organizational resources.

As the dataset has been created using resources available online, we do not have the explicit consent of the people depicted in the images. For this reason, the dataset created is not available to the public and it will remain private.

Another important limitation relates to the *ethnicity* of the people represented. We carefully constructed the dataset to include people from different backgrounds, so that a variety of ethnicities are included in the dataset. However, given the difficulties in finding images for the EG, we didn't aim at an equal distribution of different ethnicities, leaving this aspect to future work.

6 Conclusions

The goal of this study was to understand whether people with Down syndrome may experience problems when their facial image is automatically classified. By focusing on this specific group of vulnerable people, we identified a gap in the literature on bias in facial analysis systems and further contributed to the investigation of the inherent limitations of this technology.

To achieve our goal, we created a test set by collecting facial images already available on the web. We collected 400 images, 200 faces of people with Down syndrome (experimental group, EG) and 200 faces of people without the syndrome (control group, CG). We then compared the performance of two commercial face recognition tools: Amazon Rekognition (AWSR) and ClarifAI (CLAI). Overall, the findings indicated that the tools demonstrated poorer performance with the experimental group. We also found that: i) the gender prediction showed a higher error rate towards the Down male sample; ii) people with Down syndrome were assigned a younger age in relation to their real age; iii) the labels assigned to the experimental group reflected the same gender stereotypes observed in the labels of the control group, in certain cases with a higher frequency.

Involving the most vulnerable populations in the design of facial analysis systems can reduce their operational bias. Improving the transparency of documentation could also allow for better external scrutiny. However, we should question the technology itself and its implementation. Predicting sensitive characteristics, such as gender and age, as well as tagging a person's face with sensitive labels, such as aeshtetic and education, could have significant consequences for the lives of those people. Regardless of the achievable level of accuracy, this technological advancement may not be socially acceptable. This research sheds light on the structural limitations of facial analysis systems and contributes to this ongoing debate.

7 Future Work

The way in which the dataset was constructed was, for the time being, the most feasible way of investigating the research questions. One of the first improvements in the construction of the dataset is to involve people from the selected communities and ask them to take photographs of themselves, thus obtaining their consent and some valuable information. This could be a valid solution to some limitations mentioned above, such as knowing the exact age of each person and getting their explicit consent to be part of the research. Furthermore, some problems encountered during the process, such as pose, lighting and quality, can be solved by using appropriate cameras and rules for taking photographs.

An improvement concerns ethnicity. An idea to construct an equally balanced dataset can be inspired by the *Pilot Parliaments Benchmark* dataset [2].

In terms of models, it would be interesting to increase the number of models studied. This will enable obtaining a more comprehensive overview of the performance of FASs concerning underrepresented groups.

Further progress could be made on active measures to reduce discrimination against under-represented groups (such as people with Down syndrome). Since the subgroup performance issues that lead to dangerous discrimination stem most probably from under-representation and unbalanced data, it would be interesting to explore a way to measure data characteristics (such as balance) and provide ad hoc designed labels that provide ethically relevant information. Such a tool could be integrated into the AI pipeline to allow developers to be aware of the data issues and take into consideration meaningful countermeasures. This could be also useful if it is used to publish and disseminate relevant information related to public datasets that are widely used in the AI community, such as those used or mentioned in this article.

Acknowledgements. This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered

responsible for them.

References

1. Agbolade, O., Nazri, A., Yaakob, R., Ghani, A.A., Cheah, Y.K.: Down Syndrome Face Recognition: A Review. Symmetry **12**(7), 1182 (Jul 2020). https://doi.org/10.3390/sym12071182
2. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. pp. 77–91. PMLR (Jan 2018), https://proceedings.mlr.press/v81/buolamwini18a.html
3. Crawford, K.: The Atlas of AI. Yale University Press (2021)
4. European Commission: Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (2021), https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:52021PC0206
5. European Parliament, Directorate-General for Parliamentary Research Services, Madiega, T., Mildebrath, H.: Regulating Facial Recognition in the EU: In Depth Analysis. Publications Office of the European Union, LU (2021), https://data.europa.eu/doi/10.2861/140928
6. Hill, K.: Microsoft Plans to Eliminate Face Analysis Tools in Push for 'Responsible A.I.'. The New York Times (Jun 2022), https://www.nytimes.com/2022/06/21/technology/microsoft-facial-recognition.html
7. Kazemi, M., Salehi, M., Kheirollahi, M.: Down Syndrome: Current Status, Challenges and Future Perspectives. International Journal of Molecular and Cellular Medicine **5**(3), 125–133 (2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5125364/
8. Keyes, O.: The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW), 88:1–88:22 (Nov 2018). https://doi.org/10.1145/3274357
9. Klare, B.F., Burge, M.J., Klontz, J.C., Vorder Bruegge, R.W., Jain, A.K.: Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security **7**(6), 1789–1801 (Dec 2012). https://doi.org/10.1109/TIFS.2012.2214212
10. Krishna, A.: IBM CEO's Letter to Congress on Racial Justice Reform (Dec 2019), https://www.ibm.com/policy/facial-recognition-sunset-racial-justice-reforms/
11. Melendez, S.: Uber driver troubles raise concerns about transgender face recognition. Fast Company (Aug 2018), https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers
12. Muthukumar, V., Pedapati, T., Ratha, N., Sattigeri, P., Wu, C.W., Kingsbury, B., Kumar, A., Thomas, S., Mojsilovic, A., Varshney, K.R.: Understanding Unequal Gender Classification Accuracy from Face Images (Nov 2018). https://doi.org/10.48550/arXiv.1812.00099
13. Paredes, N., Caicedo-Bravo, E.F., Bacca, B., Olmedo, G.: Emotion Recognition of Down Syndrome People Based on the Evaluation of Artificial Intelligence and Statistical Analysis Methods. Symmetry **14**(12), 2492 (Dec 2022). https://doi.org/10.3390/sym14122492

14. Parlamento Italiano: Testo Coordinato del Decreto-legge 8 ottobre 2021, n. 139, recante "Disposizioni urgenti per l'accesso alle attivita' culturali, sportive e ricreative, nonche' per l'organizzazione di pubbliche amministrazioni e in materia di protezione dei dati personali." (Oct 2021), www.gazzettaufficiale.it/eli/id/2021/12/07/21A07259/sg
15. Phillips, P.J., Beveridge, J.R., Draper, B.A., Givens, G., O'Toole, A.J., Bolme, D.S., Dunlop, J., Lui, Y.M., Sahibzada, H., Weimer, S.: An introduction to the good, the bad, & the ugly face recognition challenge problem. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 346–353 (Mar 2011). https://doi.org/10.1109/FG.2011.5771424
16. Qin, B., Liang, L., Wu, J., Quan, Q., Wang, Z., Li, D.: Automatic Identification of Down Syndrome Using Facial Images with Deep Convolutional Neural Network. Diagnostics **10**(7), 487 (Jul 2020). https://doi.org/10.3390/diagnostics10070487
17. San Francisco Board of Supervisors: Ordinance No. 107-19, Chapter 19B: Acquisition of surveillance technology (May 2019), https://codelibrary.amlegal.com/codes/san_francisco/latest/sf_admin/0-0-0-47320
18. Scheuerman, M.K., Paul, J.M., Brubaker, J.R.: How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 144:1–144:33 (Nov 2019). https://doi.org/10.1145/3359246
19. Vincent, J.: Transgender YouTubers had their videos grabbed to train facial recognition software. The Verge (Aug 2017), https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset
20. West, S.M., Whittaker, M., Crawford, K.: Discriminating Systems: Gender, Race, and Power in AI. Tech. rep., AI Now Institute (2019), https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2
21. Zhang, Z., Song, Y., Qi, H.: Age Progression/Regression by Conditional Adversarial Autoencoder (Mar 2017). https://doi.org/10.48550/arXiv.1702.08423