Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces

Jaspar Pahl*
Ines Rieger*
jaspar.pahl@iis.fraunhofer.de
ines.rieger@iis.fraunhofer.de
Fraunhofer IIS
Erlangen, Bavaria, Germany
University of Bamberg
Bamberg, Bavaria, Germany

Anna Möller
Fraunhofer IIS
Erlangen, Bavaria, Germany
University of Erlangen–Nuremberg
Erlangen, Bavaria, Germany

Thomas Wittenberg
Fraunhofer IIS
Erlangen, Bavaria, Germany
University of Erlangen–Nuremberg
Erlangen, Bavaria, Germany

Ute Schmid
University of Bamberg
Bamberg, Bavaria, Germany
Fraunhofer IIS
Erlangen, Bavaria, Germany

ABSTRACT

Nowadays, Artificial Intelligence (AI) algorithms show a strong performance for many use cases, making them desirable for real-world scenarios where the algorithms provide high-impact decisions. However, one major drawback of AI algorithms is their susceptibility to bias and resulting unfairness. This has a huge influence for their application, as they have a higher failure rate for certain subgroups. In this paper, we focus on the field of affective computing and particularly on the detection of bias for facial expressions. Depending on the deployment scenario, bias in facial expression models can have a disadvantageous impact and it is therefore essential to evaluate the bias and limitations of the model. In order to analyze the metadata distribution in affective computing datasets, we annotate several benchmark training datasets, containing both Action Units and categorical emotions, with age, gender, ethnicity, glasses, and beards. We show that there is a significantly skewed distribution, particularly for ethnicity and age. Based on this metadata annotation, we evaluate two trained state-of-the-art affective computing algorithms. Our evaluation shows that the strongest bias is in age, with the best performance for persons under 34 and a sharp decrease for older persons. Furthermore, we see an ethnicity bias with varying direction depending on the algorithm, a slight gender bias and worse performance for facial parts occluded by glasses.

*Both authors contributed equally to this research.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533159

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer vision**; • **Social and professional topics** → *Age*; *Race and ethnicity*; Gender.

KEYWORDS

affective computing, action units, categorical emotions, metadata post-annotation, bias, fairness, data evaluation, algorithm evaluation

ACM Reference Format:
Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. 2022. Female, white, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3531146.3533159

1 INTRODUCTION

Artificial Intelligence (AI) has made great progress in recent years. The relative abundance of computational power and concentrated efforts to provide easily accessible, curated training data [11, 28] has led to a surge in research on AI systems [47]. As a result, a growing number of companies are integrating AI into their products, and with progressing digitalization, this trend is likely to continue. Despite all the advantages those systems provide, AI algorithms are still suffering from major problems which hinder its real world application possibilities. One particular problem is bias in automatic systems driven by AI, which has repeatedly led to major problems in fairness [4]. This is particularly impactful when the information gathered from those systems is used to assess humans regarding their skills or intentions, since a bias in such a system can potentially influence the treatment or even the future opportunities of a person. One popular example for ethnic unfairness is images of people of color being labeled as gorillas in the 2015 Google Lens scandal. This lead to Google having to remove the 'gorilla' label in it's entirety later [45]. An example for gender bias is Amazon's AI recruiting tool systematically disadvantaging women in 2018 [9].

The technology behind automatic emotional analysis of humans is called affective computing. [38]. While affective computing can also process information like speech, body posture, or physiological measurements, this study has a focus on facial affect recognition where images or videos of human faces are processed. Image-based modalities provide a high amount of information because humans use facial expressions extensively when communicating [21], while also relying on a type of sensor that is commonly available and contact free.

In facial affective computing, there are two main approaches on how to judge the information content of a face. The first one is focused on using categorical classifications such as happiness, sadness, or fear developed in the late 1960s [15, 16]. This set of originally six emotions (later extended by another 11 categorical emotions [13]) was thought to be independent of cultural background or education and to be understood by any person. However today, the idea of universal emotions is under strong criticism because more recent research has shown that, in fact, emotion displays are at least partially trained and not uniform across all people [5]. However, since it does not require special training to annotate a dataset and the annotation process is comparably fast, the categorical emotion annotation is still very popular in affective computing to this date. The main disadvantage of categorical emotion annotation is that it is a subjective rating and the annotators are influenced by their own backgrounds and interpretations. Even when the emotion is induced in a controlled setting and the emotion displays are checked afterwards, there is still no certainty regarding the success of the emotion stimulus because subjects are not uniform in their reactions to stimuli.

This problem has led to the development of the Facial Action Coding System (FACS) [14, 17]. In a FACS annotation, a face is not scored with an interpretation of an emotion, but with discrete movements of independent muscle groups called Action Units (AUs). The FACS system has been developed in order to provide a standardized, objective description manual for facial expressions and is mainly used in psychology. Since the coding is very detailed, it requires a special training and a considerable amount of time to annotate an emotion display. A full FACS coding of a 1 minute video can take up to 30 minutes [51]. Since facial affective computing is already being applied and likely to become a critical technology in the near future, this paper looks into how much bias there is in training data and how much effect bias can have in the deployment of the algorithm regarding its fairness.

Bias in emotion recognition is a known problem and has been investigated in different types of commercial software [41]. Further research both on detection and mitigation of different types of bias in emotion recognition has been conducted [20, 30, 36]. Bias in Action Unit Detection on the other hand has not been evaluated for a wide range of datasets and algorithms. Churamani et al. [7] evaluate and mitigate gender and ethnicity bias for facial expression recognition and AU detection using the RAF-DB [29, 31] and BP4D datasets. Deuschel et al. [12] intentionally induce sample bias for gender and skin color and analyze the model bias qualitatively and quantitatively using the CK+ and Actor Study datasets. Taati et al. [43] examine algorithmic bias against cognitively impaired patients for two Action Units on a small sample size. They emphasize that most datasets contain young, healthy patients and point out the potential resulting disadvantages. To the best of our knowledge, we are the first to evaluate such a broad range of Action Unit datasets while taking age and partial facial occlusions into account as a source of bias. Our evaluation reveals age as the most prominent source of bias for Action Unit detection. In the following we list our contributions and pose our research questions (in bold):

(1 a) We post-annotated the following affective computing datasets which have a focus on Action Unit detection for missing information about metadata like age, gender, ethnicity, glasses, and beards: AffWild2, BP4D, BP4D+, CK+, DISFA, DISFA+, GFT, UNBC, and ERIK.

(1 b) We analyse the resulting meta annotation distributions in those datasets.
RQ 1: Is there a bias in the metadata of AU datasets?

(2) We evaluate the susceptibility of two modern AU detection algorithms regarding bias.
RQ 2: Are state-of-the-art AU detection algorithms prone to bias?

(3) We evaluate the same algorithms regarding susceptibility to bias in their output for categorical emotions.
RQ 3: Is Action Unit detection less prone to bias than the detection of categorical emotions?

2 DATASETS

In this work, we investigate well-known AU benchmarking datasets (e.g. CK+ or DISFA) as well as more context-specific AU data (e.g. ERIK or AffWild2). To highlight the differences and commonalities Table 1 characterizes the technical aspects of all ten FACS-coded datasets.

AffWild2 is an extension of the former AffWild dataset [23, 24, 46]. It contains ground truth labels for three different tasks, leading to three different subsets which are then further split into training, validation, and testing data. In this work, the AU validation (*AU-V*) and the AU training subset (*AU-T*) are used. As the videos are recorded in-the-wild, they differ in illumination, quality, and recording angles. BP4D-Spontaneous [49], short **BP4D**, contains not only FACS-coded image sequences but also three-dimensional data. The illumination and background of the recording scenery are constant. The extension of BP4D, namely **BP4D+** [50], further includes additional subjects and multimodal data (e.g. thermal videos or physiological parameters). The Extended Cohn-Kanade Dataset [32], also known as **CK+**, is one of the most widely-used facial expression datasets. It was released in 2000 and contains Action Unit and emotion coding. The sequence length is significantly shorter than in the other datasets. Due to providing annotations for Action Units as well as categorical emotions, CK+ is used to compare the susceptibility to bias of both representations. The Denver Intensity of Spontaneous Facial Action (**DISFA**) [35] database is a manually FACS-coded collection of image sequences. While being shown a 4-minute video clip, the subjects' reactions were recorded in front of a uniform blue background and constant illumination. **DISFA+** [34] refers to an extension to the DISFA dataset where a small group of subjects took part in acted sequences. **ERIK** is a dataset containing FACS-coded images of 13 children and hasn't been published. There are other categorical emotion datasets containing children,

Table 1: Overview of technical aspects of different FACS-coded data, e.g. the number of subjects (*subj.*), the total frames, the type of affect (spontaneous (*spont.*) or acted), and whether the AUs were coded per sequence (*seq.*) or per frame. Please note that only parts of the datasets that are FACS-coded are considered here.

Dataset	Subj.	Frames/Subj.	Frames	Setup	Affect	Action Units	Coding
AffWild2-AU-V AffWild2-AU-T	7 40	4,288-21,938 180-47,435	67,306 235,944	Wild	Spont.	1, 2, 4, 6, 12, 15, 20, 25	Frame
BP4D BP4D+	41 140	3,183-4,153 1,202-2,201	146,346 197,875	Lab	Spont.	1, 2, 4-7, 9-20, 22-24, 27-39	Frame
CK+	123	7-220	10,734	Lab	Acted	1, 2, 4-7, 9-18, 20-31, 34, 38, 39, 41-46	Seq.
DISFA DISFA+	27 9	4,845 4,063-8,697	130,815 57,668	Lab	Spont. Acted	1, 2, 4-6, 9, 12, 15, 17, 20, 25, 26	Frame
ERIK	13	70-342	2,700	Lab	Acted	1, 2, 4-7, 9-18, 20, 22, 24-38, 43, 45, 51-58, 61-64	Frame
GFT	96	1,800	172,800	Lab	Spont.	1, 2, 4-7, 9-12, 14, 15, 17-19, 22-24, 28	Frame
UNBC	25	518-3,592	48,398	Lab	Spont.	4, 6, 7, 9, 10, 12, 15, 20, 25-27, 43, 50	Frame

e.g. EmoReact [37], MMDB [40], and Dartmouth [8], but large FACS-coded datasets for children are not publicly available at this point. Therefore, ERIK is used to examine how the models perform on children. Apart from high resolution and good illumination conditions, the FACS-coding was performed frame-wise, which leads to accurate information with little label-noise, therefore the technical quality is above average. The Sayette Group Formation Task Spontaneous Facial Expression Database [18], also referred to as **GFT**, shows 96 participants in social interaction in groups of three. Some of the subjects drink alcoholic beverages. Although the videos are recorded in a lab, the setup is not as controlled as for example in BP4D, and sometimes the subject's faces are occluded by glasses or other subjects. The UNBC-McMaster Shoulder Pain Expression Archive Database [33], short **UNBC**, contains video sequences of subjects that suffer from shoulder pain. The recording setting is controlled and illumination and background are similar in all sessions. It was published in 2011. It is particularly interesting since the subjects are significantly older than in other datasets. However, due to low quality images, the dataset is challenging.

3 SUBJECT META DATA

3.1 Annotation Process

The datasets presented in Section 2 have been published with different meta information regarding their subjects. Table 2 shows the details of the ground truth metadata annotation for the respective properties, which we considered in our post-annotation. A *per subject* annotation means that the ground truth information is published, but often only a distribution (denoted as *dist.*) or *range* for the cumulative data is noted in the publication. GFT is the only dataset with subject-wise ground truth information on gender, ethnicity, age, and glasses. Overall, few meta information about the appearance of subjects is published in the datasets.

To obtain complete subject-wise meta information, four independent coders estimated the missing properties retrospectively. Therefore, videos or images of each subject were watched and their gender (female, male), age (as integer), and ethnicity (African-American, Asian, Euro-American, Hispanic) was estimated based on their appearance. This set of ethnicities was used because it is commonly referred to in the papers published with the used datasets. Furthermore, the information whether a person wears glasses or has a beard was coded by one and checked by all other raters. In case of disagreement the subject was reviewed, as these attributes are observed rather than estimated. Whenever a dataset provided further information concerning the meta information, e.g. the distribution of ethnicities or an age range, these additional information were taken into consideration during the estimation process.

3.2 Annotation Quality

To rate the inter-rater-reliability the codings are evaluated with Krippendorff's α [27] using PyPi's implementation[1]. Krippendorff's α is defined as

$$\alpha = 1 - \frac{D_o}{D_e},$$

with D_o being the observed disagreement and D_e being the disagreement expected in values of the same interval assigned by chance [26]. Krippenddorff's α ranges within $[-1, 1]$, which represents systematic disagreement to systematic agreement. If α is 0, the assigned values are statistically not related. The minimum

[1] https://pypi.org/project/krippendorff/, retrieved 11/01/2021

Table 2: Subject meta information published with the respective datasets.

Attribute	AffWild2-AU	BP4D	BP4D+	CK+	DISFA	DISFA+	ERIK	GFT	UNBC
Gender	dist.	per subject	per subject	dist.	dist.	-	(forenames)	per subject	dist.
Ethnicity	-	dist.	dist.	dist.	dist.	-	-	per subject	-
Age	-	range	range	range	range	range	-	per subject	-

score required for reliability inevitably depends on the intended application, however in this context $\alpha \geq 0.8$ indicates good reliability and $\alpha \geq 0.667$ is sufficient for tentative conlusions [25]. A distance metric is required to measure the levels of agreement. To retrieve the scores presented in Table 3, the interval metric is used for age, whereas nominal distance is determined for gender and ethnicity. The **gender** estimations are the most reliable. Except for one subject in ERIK, all coders perfectly agreed for all subjects. Therefore, the values can directly be assigned to the subjects. For **ethnicity**, the scores range above 0.667 for all datasets apart from DISFA and DISFA+, which are slightly lower, as well as ERIK. The latter contains subjects that are mostly estimated as Euro-Americans or Hispanics, two ethnicities that seem to be difficult to distinguish visually, which serves as a potential explanation for the low score. Nevertheless, the overall score for the combined data of $\alpha = 0.75$ justifies the use. The scores for **age** estimations are comparably low for each dataset. However, the overall score of $\alpha = 0.85$ indicates high agreement. This disambiguity is caused by the nature of α: the disagreement is evaluated within the respective minimal and maximal value, which is narrow in a majority of the datasets. For a narrow age range the disagreement expected by chance D_e is smaller, therefore α decreases for constant D_o. The overall score relates to the range from global minimum to maximum age across datasets. Therefore, the overall score describes the actual process of assigning an age from all reasonable possibilities better and rates the age estimation for the entire data with very good reliability. For age, the averaged value of all coders is assigned to the subjects. For the subsequent evaluation of the algorithms, the ages are summarized into age groups of ten years width. The resulting data is explicitly considered estimative, as it was judged solely on appearance and the individuals could not be asked how they identify due to anonymization or missing contact data.

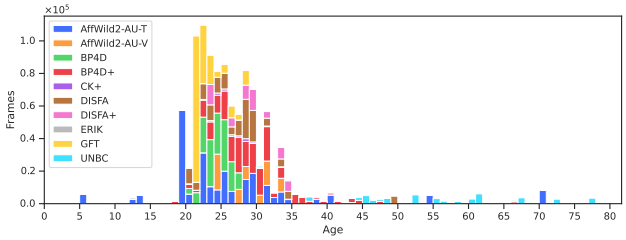

Figure 1: Age distribution of all combined datasets, with hue for dataset.

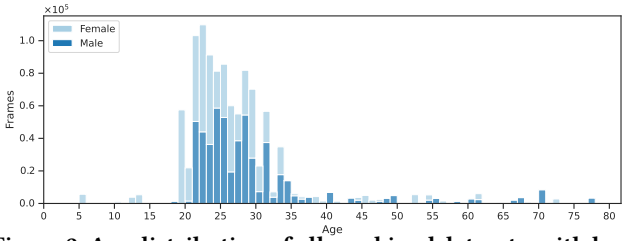

Figure 2: Age distribution of all combined datasets, with hue for gender.

3.3 Distribution of Properties

The combined Action Unit data contains 1,070,585 frames showing 526 individuals. These subjects are of different age (Figure 1 and Figure 2), gender (Figure 2 and Figure 3), and ethnicity (Figure 3). **Age** ranges from 5 to 78 years, with mean and median being 27.37 and 25 years, respectively. However, 70.97% of all frames show subjects between 20 and 30 years. Mainly ERIK and UNBC contribute to a wider range, the first containing only children and the second mostly subjects older than 35 years (see Figure 6 in the appendix for more detailed dataset-wise information on age distribution). Figure 2 shows the age distribution with hue gender, revealing that most of the subjects younger than 22 years are female, and subjects over 34 years tend to be male. The peak at 19 years is mainly caused by a single long video from AffWild2-AU-T that shows a hispanic female (see Figure 7 in the appendix for more detailed dataset-wise information on gender, ethnicity, and age). **Gender** is distributed almost uniformly, with 50.64% of frames showing females and 49.36% containing males. Out of the four **ethnicities**, Euro-Americans are present in 62.08% of frames, Asians in 15.97%, African-Americans in 11.13%, and Hispanics in 10.65%. Within the ethnicity groups, gender is not as well-balanced as in the overall data, with less female African-Americans, but more female Asians and Hispanics. There is one subject in GFT whose ground-truth ethnicity is 'Other'. Furthermore, 11.63% of frames show subjects with **glasses**, and 30.03% of males have a **beard** (see Figure 3).

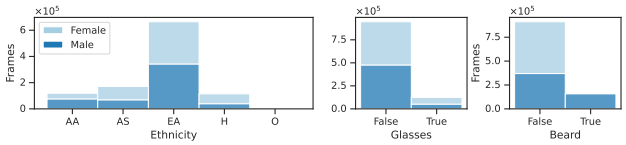

**Figure 3: Gender, ethnicity, beard, and glasses in all combined datasets. (*AA*: African-American, *AS*: Asian, *EA*: Euro-American,
HI: Hispanic, *O*: Other).**

Table 3: Krippendorff's α of our post-annotations for the respective attribute and dataset, as well as for the combined data.

Attribute	AffWild2-AU	BP4D	BP4D+	CK+	DISFA	DISFA+	ERIK	UNBC	**Combined Data**
Gender	1	-	-	1	1	1	0.95	1	**0.99**
Ethnicity	0.80	0.82	0.68	0.75	0.64	0.66	0.36	0.74	**0.75**
Age	0.92	0.34	0.57	0.46	0.66	0.35	0.39	0.71	**0.85**
Mean	**0.91**	**0.72**	**0.75**	**0.74**	**0.77**	**0.67**	**0.57**	**0.82**	**0.87**

4 MODELS

The distribution of the properties of commonly used Action Unit datasets (examined in Section 3) enforces the hypothesis that algorithms trained on this data for affect recognition are susceptible to bias. Therefore, the underlying data and functionality of a state-of-the-art model and an open-source facial analysis toolkit are presented in the following.

Algorithms have been chosen with regard to the capability to score categorical emotion and Action Units at the same time, performance in doing so, availability of the code, and being in active deployment (particularly OpenFace).

4.1 NISL2020

NISL2020 [10] is a model that produced excellent results in a competition called "Affective Behavior Analysis in-the-wild" at the 15th IEEE International Conference on Automatic Face and Gesture Recognition in 2020 [22]. It is a multi-task model for Valence/Arousal (VA) estimation, AU detection, and expression classification. The main focus of this work is on the AU detection; in this category of the competition, NISL2020 was ranked second. It was chosen for this work for its out-of-the-box functionality and reproducibility: Preprocessing algorithms as well as trained models are provided.[2]

NISL2020 explores two different approaches: The first is based on a Convolutional Neural Networks (CNNs) with a ResNet50 backbone [19], the second approach uses the ResNet50 combined with a Recurrent Neural Networks (CNN-RNNs) where Gated Recurrent Unit (GRU) layers [6] are added to captivate temporal correlations. For face extraction Deng et al. [10] use the existing preprocessing algorithm Multitask Cascaded Convolutional Network (MTCNN) [48] that detects, aligns, and crops a face. Only faces that are successfully detected are processed by the networks. They propose a student-teacher approach to tackle the challenge of incompletely labeled data by using a supervised trained teacher to propose soft labels for unlabeled data when training the student. As the NISL2020 CNN-RNN approach with a sequence length of 32 frames is reported to have the best performance, this approach is used on all datasets except for CK+. This database contains sequences that are too short for this approach, therefore, for comparability, the CNN approach is used for the entire dataset. NISL2020 returns the predictions of the teacher model and five student models. The students' results are averaged and binarized using individual thresholds per Action Unit. This merged binary output is used as the final model prediction.

The authors of NISL2020 used the datasets AffWild2 and DISFA (see Section 2) to train and validate their Action Unit detection. Please note that we can only analyze the training data for the AU detection, but since NISL2020 is a multi-task network the distribution of the training datasets for VA and categorical emotions potentially also have an influence. Both DISFA and AffWild2 lacked subject meta information in the original datasets, so it was post-annotated as described in Section 3. The Action Unit training data for NISL2020 consists of 67 individuals with 180-47,435 frames per individual. This results in 366,758 total frames. Of these frames, 53.46% display female subjects, opposed to 46.54% that show males. Figure 4 shows how different ethnicities are distributed. More than half of the frames contain Euro-American subjects. Furthermore, the gender imbalance within the African-American class is remarkable. Although 8.59% of all samples show African-American subjects, only 1.05% of the total frames show African-American females. Figure 5 shows that the underlying training data of NISL2020 tends to represent subjects between 15 and 34 years significantly better than all other subjects. Whereas subjects between 15 and 24 years and 25 and 34 years are shown in 42.82% and 45.82% of the images, all other age groups are in less than 4% of the images respectively. The age mean of all displayed faces is 26.47 years with a slightly lower median age of 25 years. The minimum age is 5 years, the maximum age 70 years. Furthermore it is denoted that 17.77% of the total frames show subjects with beards, which are 38.19% of the male subjects. Glasses are worn in 23.58% of all frames.

4.2 OpenFace

OpenFace 2.0 [2] is an open source toolkit[3] with different facial analysis functionalities, like eye gaze tracking or facial landmark detection. It also provides AU detection algorithms [1]. OpenFace has been chosen as the second model in this study due to being the most popular and accessible open source AU detection.

OpenFace detects 18 Action Units, including all 8 that are recognized by NISL2020, and is capable of scoring activity and intensity of displayed Action Units. In our study, the binary activity values are used for better comparability within the meta-groups. According to the repository[4] that provides the resources for release 2.2.0, seven different datasets have been used for training the AU detection model: BP4D [49], CK+ [32], DISFA [35], UNBC [33], and three other datasets that are not part of our examinations [3, 42, 44]. Thus, the metadata distribution concerning age, ethnicity, and gender of these datasets is not analyzed here. Furthermore, this also implies that these training datasets need to be excluded from testing OpenFace.

5 RESULTS

5.1 Dataset-wise Performance Evaluation

As outlined in Section 2, the testing datasets differ in many aspects, e.g. recording setup, quality, but also the displayed subjects. Therefore, the performance of the models needs to be judged with regard

[2]https://github.com/wtomin/Multitask-Emotion-Recognition-with-Incomplete-Labels, retrieved 05/02/2021.

[3]https://github.com/TadasBaltrusaitis/OpenFace, retrieved 05/02/2021.

[4]https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units, retrieved 02/02/2022.

Figure 4: Distribution of ethnicity and gender in the Action Unit training data of NISL2020 shown as the number of subjects (denoted as *Subj.*), frames per ethnicity (denoted as *Ethn.*), and gender (denoted as *Gend.*). *AA* stands for African-American, *AS* for Asian, *EA* for European-American, and *HI* for Hispanic. *F* and *M* are short for female and male. Highest numbers are in bold. The bar charts on the right show the distribution of glasses and beards.

Ethn.	Subj.	Frames	Gend.	Subj.	Frames
AA	9	31,515	F	2	3,850
			M	7	27,665
AS	7	38,372	F	4	18,861
			M	3	19,511
EA	**44**	**214,127**	F	**23**	**115,375**
			M	21	98,752
HI	7	82,744	F	4	57,970
			M	3	24,774

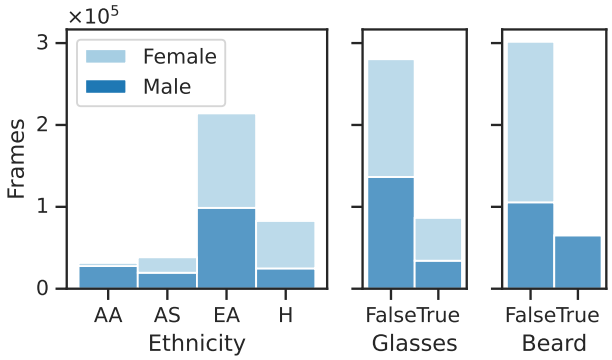

Figure 5: Distribution of age in the Action Unit training data of NISL2020 shown as the number of subjects (denoted as *Subj.*) and frames per age group. Highest numbers are in bold. The bar chart on the right shows the age distribution separated by datasets.

Age	Subj.	Frames
5-14y	5	13,890
15-24y	22	157,058
25-34y	**31**	**168,052**
35-44y	4	8,840
45-54y	3	10,084
55-64y	1	415
65-74y	1	8,419

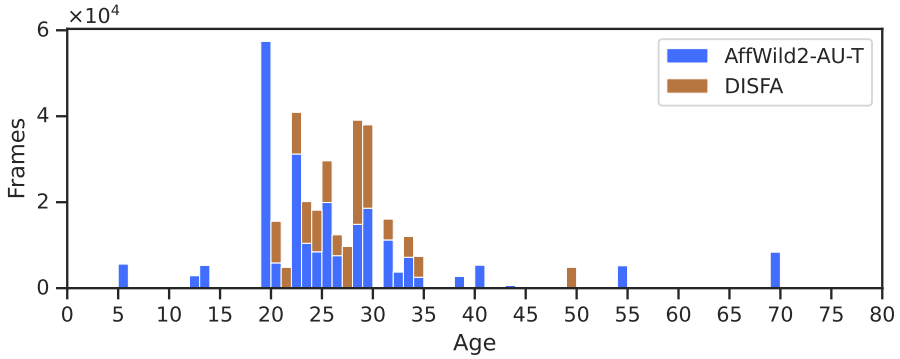

Table 4: Dataset-wise performance evaluation per model. For each model, the minimum and maximum macro F1-score is bold. The numbers of *Subjects* and *Frames* refer to testing data, not training data.

		NISL2020		NISL2020 cp		OpenFace	
Dataset	Subjects	Frames	F1-Macro	Frames	F1-Macro	Frames	F1-Macro
AffWild2-AU-T	40	-	-	-	-	213,077	**0.332**
AffWild2-AU-V	7	52,446	0.383	67,306	**0.204**	67,191	0.420
BP4D	41	121,016	0.657	146,346	0.512	-	-
BP4D+	140	182,147	0.602	197,875	0.544	-	-
CK+	123	10,711	0.639	10,734	0.636	-	-
ERIK	13	2,699	**0.723**	2,700	**0.723**	2,700	**0.649**
GFT	96	160,555	0.508	172,800	0.438	172,800	0.399
UNBC	25	47,397	**0.271**	48,398	0.253	-	-

to the dataset. Table 4 shows the overall performance of the models on all available testing datasets separately. As OpenFace was trained on BP4D, CK+, and UNBC, and due to BP4D+'s similarity to BP4D, these datasets are excluded when evaluating OpenFace. For the same reason, the AffWild2 AU Training subset is used to evaluate OpenFace, but excluded for NISL2020. To ensure equal conditions, only Action Unit 4 (*brow lowerer*), 6 (*cheek raiser*), and 12 (*lip corner puller*) are considered, due to these being the Action Units that are predicted by both models and annotated as ground truth data in all used datasets.

We measure bias by comparing the F1-score

$$F1 = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

across groups, where *tp* are the true positives, *fp* the false positives, and *fn* the false negatives. The F1-score ranges in [0,1], with 1 being the score for a perfect classifier. The macro F1-score is a non-weighted mean over the F1-scores of the classes. It is commonly used for AU detection, as this metric is suited for multi-class problems with a high ratio of negative samples.

Conspicuously, MTCNN, the pre-processing algorithm used with NISL2020, only detects faces in 89.3% of the frames, although especially the lab-recorded sequences (see Section 2) contain faces in every frame. Logically, the faces that are not detected are not classified concerning Action Units. This effect is further examined in Section 5.3. To separate this error from the actual Action Unit detection, the output of NISL2020 is scored in two different ways: The results that are denoted with *NISL2020* contain the scores for all faces that were actually detected. For NISL2020, the frames where MTCNN failed are excluded to ensure that a potential resulting bias of the overall system is caused by the Action Unit detection part, not the preprocessing. However, as the preprocessing is part of the model, and the error in MTCNN would definitely result in an overall error in a use-case scenario, the faces that are not detected are included in the results called *NISL2020 cp*. For NISL2020 cp, the missing frames were padded using the contrary of the ground truth each. This fully takes the system failure into account and increases the number of false predictions without unintentionally adding correct values. In comparison, simply evaluating all AUs in frames which have not been found by the face detector as negative annotations by the AU detection algorithm would increase its performance since the majority of ground truth annotations in AU datasets are negative.

NISL2020 performs best on ERIK and worst on UNBC. The datasets BP4D, BP4D+, CK+, and GFT have a higher score than the macro F1-score for the combined data (0.586). In comparison, the contrary-padded predictions of NISL2020 cp show a decreased performance for AffWild2, BP4D, BP4D+, and GFT, whereas the predictions for ERIK and UNBC are not scored significantly worse. This aligns with the ratio of recognized faces: AffWild2 (77,92%), BP4D (82,35%), BP4D+ (92.1%), GFT (92,91%), ERIK (99.96%), CK+ (99.93%), and UNBC (97.93%). Similarly, OpenFace predicts best for ERIK, and significantly worse for AffWild2 and GFT.

5.2 OpenFace

OpenFace is evaluated on five Action Units (see e.g. Table 5) which are selected because they are contained in both the ground truth of the used datasets and the model's annotation. The **ethnicity-wise** comparison of OpenFace's performance shows the highest reliability for African-American subjects, with a slightly lower macro F1-score for European-American subjects. This tendency remains in AU-wise examination. The lowest-scored groups are Hispanics and Others. These groups, however, contain only seven and one subjects respectively, compared to 121 Euro-Americans. Therefore, the result may not be abstractable. The **gender-wise** comparison shows a difference of 0.05 for the macro F1-score, slightly in favor of male subjects. This tendency is reflected in the AU-wise evaluation as well, with the highest difference between male and female being 0.18 for AU 1, the inner brow raiser. The recognition of AU 12, the lip corner puller, is slighlty more successful in females. The number of subjects and frames in both groups is comparably balanced. Evaluating OpenFace on groups separated according to **age** results in the following observations: Regarding the macro F1-score, the subjects aged between 5-14 years are rated most successfully. It has to be noted that nearly all of the subjects in this age category come from the ERIK dataset, which has the highest dataset quality by a large margin which can be seen in Table 4. The difference to the two subsequent age groups is 0.021 and 0.016 respectively, followed by a stronger decrease of the macro F1-score with increasing age. This relation is not fully consistent for each single AU. However, it needs to be denoted that the results are less reliable for the age groups which have fewer frames in the testing data. Separating the male subjects based on whether they are wearing a **beard** does not yield any insightful results. All AUs except for AU 12, the lip corner puller, are in the eye area, and the difference for AU 12 is lower than the difference of the macro F1-scores and most of the seemingly non-beard-related AUs. OpenFace's macro F1-score and the individual scores of AU 1, 2, 6, and 12 decrease for subjects with **glasses** by an average of 0.1026, the score for AU 4 increases slightly by 0.01. The decrease for AU 6, the cheek raiser, is 0.176. As all of the AUs, except for AU 12, are in the upper face area, this indicates that recognition can be hindered by glasses.

5.3 NISL2020

The **ethnicity-wise** evaluation in Table 6 shows no clear differences between the ethnicities of the NISL2020 output, with the macro F1-scores differing by 0.099 at most. However, comparing the clean NISL2020 and the contrary-padded NISL2020 cp version shows that only 48,8% of African-American faces are detected, compared to 95.11% for all other subjects. Consequently, the contrary-padded predictions show a significant difference of 0.437 between the macro F1-score of African-American and Asian subjects. The result for African-American subjects being the lowest aligns with the training data, where the African-American subjects are the least represented (see Figure 4). Further dataset-wise examinations indicate that a majority of undetected faces originated from BP4D, a dataset with a dark-blue background and soft illumination. For African-American females, only 19.98% of faces are detected, for African-American males only 6.79%, compared to more than 98% for all other ethnicities. Contrarily, within GFT, a dataset with a bright background and better illumination, 85.8% of African-American faces are detected, compared to 93.8% for Euro-Americans. This leads to the assumption that the low contrast influences the performance. However, contrast enhancing using CLAHE [39] did not increase the ratio of detected faces sufficiently.

The **gender-wise** comparison in Table 7 shows that AU 4 is better recognized in males, and AU 6 and 12 in females, which also results in a higher macro F1-score for females. The better macro results when evaluating on female subjects are in sync with the training data, since there are more female subjects present (see Figure 4). The contrary-padded output is excluded in this and the following evaluations, because it does not contribute any additional insights. Evaluating the models on groups separated according to their **age**

Table 5: F1-scores for attribute-wise evaluation of OpenFace on its testing data (AffWild2-AU-T, AffWild2-AU-V, ERIK, and GFT). For each attribute and Action Unit, the maximum and minimum F1-scores are bold.

Attribute	Value	Subjects	Frames	AU 1	AU 2	AU 4	AU 6	AU 12	Macro
Ethnicity	African-American	20	54,988	**0.657**	**0.197**	**0.415**	**0.529**	0.444	**0.448**
	Asian	7	28,894	0.376	0.081	0.365	0.354	**0.330**	0.301
	Euro-American	121	295,110	0.410	0.146	0.319	0.481	**0.516**	0.374
	Hispanic	7	74,976	0.461	0.078	0.381	**0.042**	0.450	0.282
	Other	1	1,800	**0.102**	**0.042**	**0.316**	0.341	0.463	**0.253**
Gender	Female	73	214,916	0.367	0.136	0.319	0.421	**0.508**	0.350
	Male	83	240,852	**0.547**	**0.139**	**0.371**	**0.471**	0.473	**0.400**
Age	5-14y	18	14,661	0.363	0.109	**0.498**	**0.514**	0.475	**0.392**
	15-24y	96	263,906	0.387	**0.186**	0.287	0.470	**0.524**	0.371
	25-34y	34	155,469	0.546	0.077	0.421	0.400	0.438	0.376
	35-44y	4	8,836	0.470	0.000	0.304	0.448	0.323	0.309
	45-54y	2	5,237	**0.768**	0.001	0.216	0.174	0.340	0.300
	55-64y	1	414	**0.048**	0.000	**0.000**	0.000	**0.656**	0.141
	65-74y	1	7,245	0.184	0.000	0.079	0.000	0.031	**0.059**
Beard	No	57	164,292	0.532	**0.153**	0.335	0.428	**0.481**	0.386
	Yes	26	76,560	**0.571**	0.113	**0.436**	**0.550**	0.455	**0.425**
Glasses	No	136	350,115	**0.478**	**0.164**	0.340	**0.485**	**0.503**	**0.394**
	Yes	20	105,653	0.389	0.053	**0.350**	0.309	0.450	0.310

Table 6: F1-scores for ethnicity-wise evaluation of *NISL2020* and the contrary-padded output *NISL2020 cp*. For each Action Unit, the minimum and maximum F1-scores are bold.

		NISL2020					NISL2020 cp				
Ethnicity	Subj.	Frames	AU 4	AU 6	AU 12	Macro	Frames	AU 4	AU 6	AU 12	Macro
African-American	53	39,611	**0.283**	0.619	**0.722**	**0.541**	81,164	**0.058**	0.277	**0.217**	**0.184**
Asian	73	124,477	0.375	**0.765**	0.780	**0.640**	126,921	**0.351**	**0.749**	**0.762**	**0.621**
Euro-American	295	380,449	0.286	0.683	0.750	0.573	404,960	0.240	0.633	0.684	0.519
Hispanic	23	30,928	0.300	0.711	0.750	0.587	31,314	0.292	0.702	0.739	0.577
Other	1	1,506	**0.661**	**0.354**	**0.875**	0.630	1,800	0.303	**0.256**	0.450	0.336

shows a significantly better macro F1-score in ages ranging from 5-14 years, especially caused by a high individual score for AU 4, the brow lowerer. As already mentioned in OpenFace's results, this needs to be interpreted with regard to the 5-14 years age group consisting of children from the ERIK dataset mostly which has the highest performance. The performance decreases with age; the macro F1-score for 45-54 year old subjects is less than half of the score for 25-34 year olds. On subjects above 74 years, the performance is the worst. When remembering the training data in Figure 5, most subjects were also in the range of 25-34 years. However, it needs to be noted that the reliability of the results needs to be judged with the different numbers of frames in the test set in mind (nearly 100 times more frames in age range 15-24 years compared to in 5-14 years). The scores for all male subjects regarding **beards** do not show an obvious pattern, the macro F1-scores barely differ. Separating the subjects based on whether they are wearing **glasses** shows an interesting effect for the NISL2020 predictions. Whereas the score for AU 12, the lip corner puller, decreases only slightly, the F1-scores for AU 4 (brow lowerer) and 6 (cheek raiser), both having effects around the eye area, decrease significantly by 0.218 and 0.133 respectively.

5.4 Categorical Emotions and Action Units

The dataset CK+ contains not only Action Unit ground truth information, but also the categorical emotion annotation as described in Section 2. It can therefore be used to compare NISL2020's predictions and potential bias in Action Unit and emotion recognition. Only the subset of CK+ that contains annotations for both facial expression representations was used in order to ensure equal conditions in the comparison. Furthermore, NISL2020's CNN approach is required when using NISL, because there are sequences in CK+ that are too short for the CNN-RNN.

Table 8 and Table 9 show the values for comparing the performance **gender-wise**. Women's emotions are scored slightly more reliably than men's, whereas the AU predictions are slightly more reliable

Table 7: F1-scores for the attribute-wise evaluation of NISL2020 on its testing datasets (Affwild2-AU-V, BP4D, BP4D+, CK+, ERIK, GFT, and UNBC). For each attribute and Action Unit, the maximum and minimum F1-scores are bold.

Attribute	Values	Subjects	Frames	AU 4	AU 6	AU 12	Macro
Gender	Female	254	291,562	0.294	**0.753**	**0.796**	**0.614**
	Male	91	1 285,409	**0.312**	0.632	0.696	0.547
Age	5-14y	13	2,699	**0.801**	0.670	0.698	**0.723**
	15-24y	166	260,844	0.404	0.680	**0.768**	0.617
	25-34y	219	246,622	0.301	0.736	0.763	0.600
	35-44y	23	23,870	0.194	**0.780**	0.742	0.572
	45-54y	9	16,717	0.116	0.437	**0.335**	0.296
	55-64y	9	14,828	0.027	0.291	0.354	0.224
	65-74y	5	8,608	0.117	0.515	0.440	0.357
	75-84y	1	3,360	**0.000**	**0.128**	0.525	**0.218**
Beard	No	142	210,783	**0.318**	0.628	**0.700**	**0.549**
	Yes	49	74,734	0.298	**0.646**	0.687	0.544
Glasses	No	428	549,831	**0.309**	**0.703**	**0.759**	**0.590**
	Yes	18	27,140	0.091	0.570	0.705	0.455

for men than for women. Comparing the two emotion representations **ethnicity-wise** yields slightly different tendencies for both taxonomies. The macro F1-score for AUs is almost equal for African-American, Asian, and Euro-American subjects, whereas the predictions for Hispanic subjects are scored significantly lower (see Table 11). The latter effect also persists in the emotion annotation, but the first three ethnicities are rated differently, with emotions in African-American subjects being recognized the most successfully (see Table 10). The difference of the scores for African-American and Hispanic subjects differs by less than a tenth in emotion classification and AU detection. Again, it needs to be noted that the number of test frames differs by up to factor 35 between groups and the overall number of samples is small in some groups, which reduces reliability of the results in those groups.

6 DISCUSSION

The results in Section 5 can be used to judge if current affective computing algorithms are prone to bias and provide a guideline for such evaluations. The focus of this paper, however, is not on comparing the two models OpenFace and NISL2020. Those algorithms have been trained on different training data and we have to test them on different testing data in order to avoid testing on training data. Table 4 shows clearly why comparisons of algorithms have to be conducted on the same test dataset in order to be meaningful.

The most visible bias in both models regards **age**: The performance for the age group 5-14 years is the highest and decreases with increasing age. While the performance difference is relatively small for the different cohorts within 5-34 years, it gets larger between the older age categories. This relates back to the training data, where we can see that throughout most analyzed datasets the age group 15-34 years is the best represented, with an overall mean age of 27.37 years (see Figure 1). Hence, training data is not sufficiently provided for subjects exceeding this age category. We notice that the youngest category (5-14 years) is not well represented in all analyzed datasets either, but the two algorithms perform best for these subjects. We hypothesize that the models show such a high performance for the youngest category because these are mainly subjects from the ERIK dataset, which has very high image and annotation quality and contains children in the 5-14 years interval exclusively. Older subjects are underrepresented in the analyzed datasets, as well, and the two algorithms accordingly display the worst Action Unit detection performance on these older categories. Overall, wrinkles and the quality of images seem to have a big influence for Action Unit detection. Regarding **ethnicity**, OpenFace performs best on African-Americans, while NISL2020 performs best on Asians. This is surprising because our annotated data clearly shows that European-American appearance is overly represented in all benchmark datasets. A **gender** bias can be found in both models, where OpenFace recognizes Action Units in male subjects better and NISL2020 in female subjects. This is in line with the NISL training dataset having more females than males. Unfortunately, we cannot test this hypothesis for OpenFace because its training data is not known exactly. Looking at Action Unit occlusion, for example by **glasses**, we can see a worse performance on these specific Action Units in both models. For example Action Unit 6, the cheek raiser effecting the eye, is detected worse by both models when the subjects wear glasses.

We also did a preliminary evaluation on bias in categorical emotion annotation compared to Action Unit annotation according to FACS by evaluating the multi-task NISL2020 model on the CK+ dataset. It can be noted that the overall performance is worse for detecting emotions than for detecting Action Units, but bias is visible concerning both gender and ethnicity. We can see that the bias for each taxonomy is different, for example Action Units are better detected in male subjects and emotions in female subjects. However, a consistent difference in bias susceptibility was not found between the annotation styles for the data and algorithms in question.

A possible influence in our bias evaluation is the correlation of the metadata categories, for example that most subjects younger than 22 years are female (see Fig. 2). This statistical relation can

Table 8: Evaluation in F1-score of NISL2020's emotion classification per gender on CK+ dataset. The maximum results are bold.

Gender	Subjects	Frames	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Macro
Female	81	3,933	**0.292**	**0.345**	0.147	0.803	**0.333**	0.722	**0.440**
Male	37	1,936	0.113	0.335	**0.173**	**0.824**	0.187	**0.743**	0.396

Table 9: Evaluation in F1-score of NISL2020's Action Unit detection per gender on CK+ dataset. The maximum results are bold.

Gender	Subjects	Frames	AU 1	AU 2	AU 4	AU 6	AU 12	AU 15	AU 20	AU 25	Macro
Female	81	3,933	0.695	0.509	0.708	**0.601**	0.748	0.310	0.161	0.779	0.564
Male	37	1,936	**0.738**	**0.656**	**0.747**	0.538	**0.785**	**0.360**	**0.225**	**0.812**	**0.608**

Table 10: F1-scores for NISL2020's emotion classification per ethnicity on CK+ dataset. The minimum and maximum results are bold.

Ethnicity	Subjects	Frames	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Macro
African-American	14	774	**0.616**	**0.242**	0.093	0.848	**0.509**	**0.504**	**0.469**
Asian	5	224	0.424	0.481	0.000	**0.935**	0.222	0.588	0.442
Euro-American	94	4,738	0.099	0.346	**0.190**	**0.798**	0.244	0.758	0.406
Hispanic	5	133	**0.000**	**0.576**	**0.000**	0.828	**0.000**	**0.852**	**0.376**

Table 11: F1-scores of NISL2020's Action Unit detection per ethnicity on CK+ dataset. The minimum and maximum results are bold.

Ethnicity	Subj.	Frames	AU 1	AU 2	AU 4	AU 6	AU 12	AU 15	AU 20	AU 25	Macro
African-American	14	774	0.691	0.494	0.799	**0.396**	**0.685**	0.486	**0.279**	**0.722**	0.569
Asian	5	224	**0.632**	**0.069**	**0.810**	**0.713**	**0.914**	**0.571**	0.000	**0.871**	**0.573**
Euro-American	94	4,738	**0.719**	**0.592**	0.693	0.593	0.756	0.246	0.155	0.798	0.569
Hispanic	5	133	0.658	0.431	**0.545**	0.649	0.833	0.000	0.000	0.731	**0.481**

influence the results since a more successful detection of Action Units in female subjects could potentially imply more reliable predictions for young subjects. We could not detect such an effect, but the existence cannot be ruled out.

Lastly, we want to emphasize the importance of being aware of the use case of a model throughout its entire life-cycle. Bias itself is not necessarily a problem as long as the model is used in a suitable environment: Using a model which detects facial expressions well in middle-aged subjects in an environment with children is not advisable. At the same time it would be perfectly suited when working with university students. However, training data usually does not favor edge cases and finding correctly biased models for these groups may proof to be difficult. Either way, for the purpose of these decisions information on performance depending on different use-case scenarios needs to be available. Therefore, in order to enable informed decisions on model applications, we want to encourage researchers and industry to test algorithms for bias and report their results.

7 CONCLUSION

With cameras getting smaller and cheaper and machine learning improving constantly, facial affective computing on larger scales is becoming a possibility in many application fields. This raises the question how fair these algorithms and how biased their underlying datasets are.

In this paper, we first annotated and investigated the distributions of age, gender, ethnicity, glasses, and beards for prominent Action Unit datasets regarding bias. **(Research Question 1:)** We found that while those distributions differ with regard to the dataset in question, there is a significant bias towards age with a strong focus of most datasets on young adults. Furthermore, there is a very strong bias regarding ethnicity with most of the subjects having a Euro-American appearance. There is a small bias in gender with varying direction depending on the dataset, combining all datasets shows a slight bias towards female subjects.

Afterwards, we evaluated two state-of-the-art algorithms for their performance in the different meta-groups. **(Research Question 2:)** There were biases towards varying directions for gender and ethnicity, which is reasonable given that the algorithms were trained

on different datasets. The strongest bias is in age, where both algorithms had problems with elderly people. Occlusions from glasses proved to be a problem for some Action Units close to the eyes, beards surprisingly did not have a significant effect.

In the last part we compared the objective AU coding style with the subjective categorical emotion coding style in regards to susceptibility to bias. **(Research Question 3:)** While the performance of the model for Action Unit detection showed a considerably higher performance than for categorical emotion detection, the difference in bias susceptibility is comparable on the dataset and algorithm in question.

These findings lead to the following conclusions: First of all, since bias susceptibility is a problem for facial affective computing algorithms and the consequences of such bias could be severe in critical areas, future affective computing datasets should contain meta-information on the subjects. If possible, they should also contain sufficient subjects of each group in at least age, gender, and ethnicity to facilitate the development of fair algorithms. In a second step, this allows researchers to test their affective computing algorithms for fairness and to develop new algorithms, which are less sensible to bias in their training data. In fact, this is what we want to inspire with this paper. In the last steps towards application, industry integrating affective computing research in their products should actively be alerted to fairness issues by research partners, and the resulting products should contain information about their reliable areas of operation regarding bias.

ACKNOWLEDGMENTS

The authors would like to thank Robert Obermeier and Jan Adelhardt for helping in preprocessing the data.

The work presented in this paper is funded by Grant No. 01IS18056A/B of BMBF ML-3 (Federal Ministry of Education and Research) (TraMeExCo), by Grant No. 405630557 of DFG (German Research Foundation) (PainFaceReader), and by Grant No. 16SV7945K of BMBF (ERIK).

REFERENCES

[1] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset Learning and Person-specific Normalisation for Automatic Action Unit Detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 06. 1–6. https://doi.org/10.1109/FG.2015.7284869

[2] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[3] Tanja Bänziger and Klaus R Scherer. 2010. Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus. *Blueprint for Affective Computing: A Sourcebook* 2010 (2010), 271–94.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in Machine Learning. *NIPS Tutorial* 1 (2017), 2017.

[5] Lisa Feldman Barrett. 2017. *How Emotions are made: The Secret Life of the Brain.* Houghton Mifflin Harcourt.

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. (2014). arXiv:1406.1078 [cs.CL]

[7] Nikhil Churamani, Ozgur Kara, and Hatice Gunes. 2021. Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition. *arXiv preprint arXiv:2103.08637* (2021).

[8] Kirsten Dalrymple, Jesse Gomez, and Brad Duchaine. 2013. The Dartmouth Database of Children's Faces: Acquisition and Validation of a New Face Stimulus Set. *PloS one* 8 (11 2013), e79131. https://doi.org/10.1371/journal.pone.0079131

[9] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. [Online; accessed 05-January-2022].

[10] Didan Deng, Zhaokang Chen, and Bertram E Shi. 2020. Multitask Emotion Recognition with Incomplete Labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. IEEE Computer Society, 828–835.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A Large-scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 248–255.

[12] Jessica Deuschel, Bettina Finzel, and Ines Rieger. 2021. Uncovering the Bias in Facial Expressions. *Kolloquium Forschende Frauen 2020 - Gender in Gesellschaft 4.0: Beiträge Bamberger Nachwuchswissenschaftlerinnen* (2021).

[13] Paul Ekman. 1999. Basic Emotions. *Handbook of Cognition and Emotion* 98, 45-60 (1999), 16.

[14] Paul Ekman, WV Friesen, and JC Hager. 2002. Facs Manual. *A Human Face* (2002).

[15] Paul Ekman and Wallace V Friesen. 1971. Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124.

[16] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. 1969. Pan-cultural Elements in Facial Displays of Emotion. *Science* 164, 3875 (1969), 86–88.

[17] E Friesen and Paul Ekman. 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Palo Alto* 3, 2 (1978), 5.

[18] Jeffrey M. Girard, Wen-Sheng Chu, László A. Jeni, and Jeffrey F. Cohn. 2017. Sayette Group Formation Task (GFT) Spontaneous Facial Expression Database. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 581–588. https://doi.org/10.1109/FG.2017.144

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[20] Ayanna Howard, Cha Zhang, and Eric Horvitz. 2017. Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*. IEEE, 1–7.

[21] Rachael E Jack and Philippe G Schyns. 2015. The Human Face as a Dynamic Tool for Social Communication. *Current Biology* 25, 14 (2015), R621–R634.

[22] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. 2020. Analysing Affective Behavior in the First ABAW 2020 Competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 794–800.

[23] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep Affect Prediction In-the-wild: Aff-wild Database and Challenge, Deep Architectures, and Beyond. *International Journal of Computer Vision* (2019), 1–23.

[24] Dimitrios Kollias and Stefanos Zafeiriou. 2019. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv preprint arXiv:1910.04855* (2019).

[25] Klaus Krippendorff. 2004. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research* 30, 3 (2004), 411–433.

[26] Klaus Krippendorff. 2011. Computing Krippendorff's Alpha - Reliability. (2011).

[27] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to its Methodology.* Sage publications.

[28] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning Multiple Layers of Features from Tiny Images. (2009).

[29] Shan Li and Weihong Deng. 2018. Reliable Crowdsourcing and Deep Locality - Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Transactions on Image Processing* 28, 1 (2018), 356–370.

[30] Shan Li and Weihong Deng. 2020. A Deeper Look at Facial Expression Dataset Bias. *IEEE Transactions on Affective Computing* (2020).

[31] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable Crowdsourcing and Deep Locality-preserving Learning for Expression Recognition in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2852–2861.

[32] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. 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*. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[33] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon, and Iain Matthews. 2011. Painful Data: The UNBC-McMaster Shoulder Pain Expression Archive Database. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 57–64. https://doi.org/10.1109/FG.2011.5771462

[34] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. 2016. Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1–8.

[35] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. 2013. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing* 4, 2 (2013), 151–160. https://doi.org/10.1109/T-AFFC.2013.4

[36] Mkhuseli Ngxande, Jules-Raymond Tapamo, and Michael Burke. 2020. Bias Remediation in Driver Drowsiness Detection Dystems using Generative Adversarial Networks. *IEEE Access* 8 (2020), 55592–55601.

[37] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis-Philippe Morency. 2016. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 137–144. https://doi.org/10.1145/2993148.2993168

[38] Rosalind Wright Picard. 1995. Affective Computing. (1995).

[39] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. 1987. Adaptive Histogram Equalization and its Variations. *Computer Vision, Graphics, and Image Processing* 39, 3 (1987), 355–368.

[40] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. 2013. Decoding Children's Social Behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3414–3421.

[41] Lauren Rhue. 2018. Racial Influence on Automated Perceptions of Emotions. *Available at SSRN 3281765* (2018).

[42] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. 2008. Bosphorus Database for 3D Face Analysis. In *European Workshop on Biometrics and Identity Management*. Springer, 47–56.

[43] Babak Taati, Shun Zhao, Ahmed B Ashraf, Azin Asgarian, M Erin Browne, Kenneth M Prkachin, Alex Mihailidis, and Thomas Hadjistavropoulos. 2019. Algorithmic Bias in Clinical Populations — Evaluating and Improving Facial Analysis Technology in Older Adults with Dementia. *IEEE Access* 7 (2019), 25527–25534.

[44] Michel Valstar, Timur Almaev, Jeffrey Girard, Gary Mckeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey Cohn. 2015. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. https://doi.org/10.1109/FG.2015.7284874

[45] James Vincent. 2018. Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai. [Online; accessed 05-January-2022].

[46] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. 2017. Aff-wild: Valence and Arousal 'In-the-wild' Challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 1980–1987.

[47] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault. 2021. The AI Index 2021 Annual Report. (03 2021).

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (Oct 2016), 1499–1503. https://doi.org/10.1109/lsp.2016.2603342

[49] Xing Zhang, Lijun Yin, Jeffrey Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey Girard. 2014. BP4D-Spontaneous: A High-resolution Spontaneous 3D Dynamic Facial Expression Database. *Image and Vision Computing* 32 (06 2014), 692–706. https://doi.org/10.1016/j.imavis.2014.06.002

[50] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3438–3446.

[51] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. 2018. Learning Facial Action Units from Web Images with Scalable Weakly Supervised Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2090–2099.

A METADATA DISTRIBUTION PER DATASET

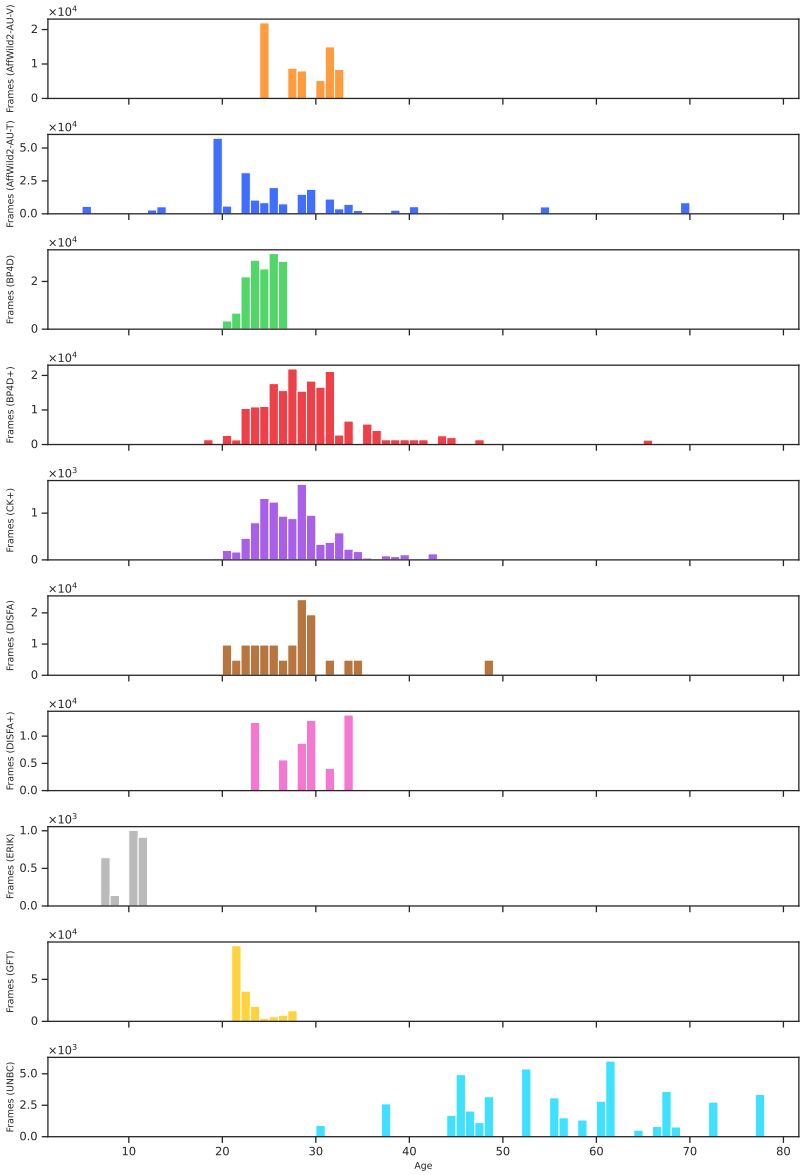

Figure 6: Age distribution of each dataset. Please note that the y-scales differ due to the varying number of frames per dataset (see Table 1).

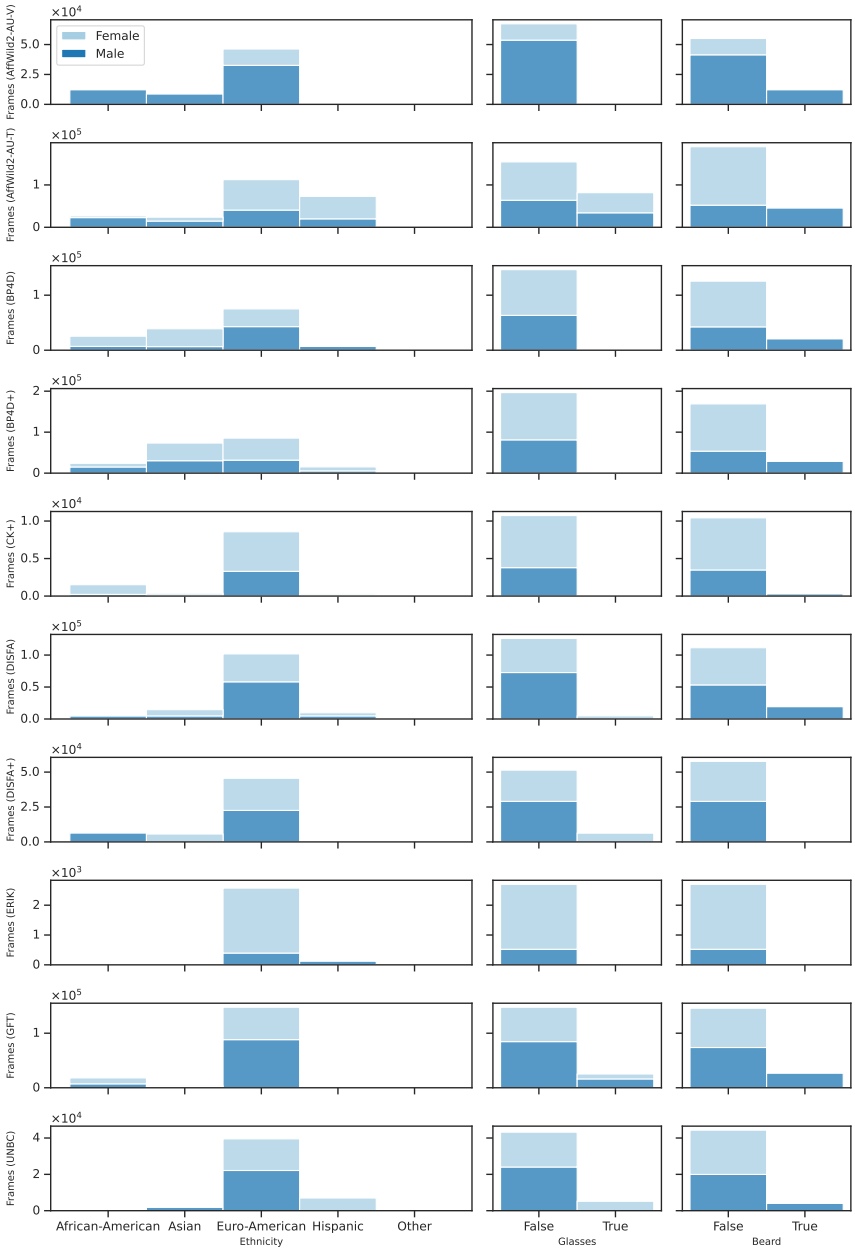

Figure 7: Distribution of ethnicity, glasses and beards within each dataset. Please note that the y-scales differ due to the varying number of frames per dataset (see Table 1).

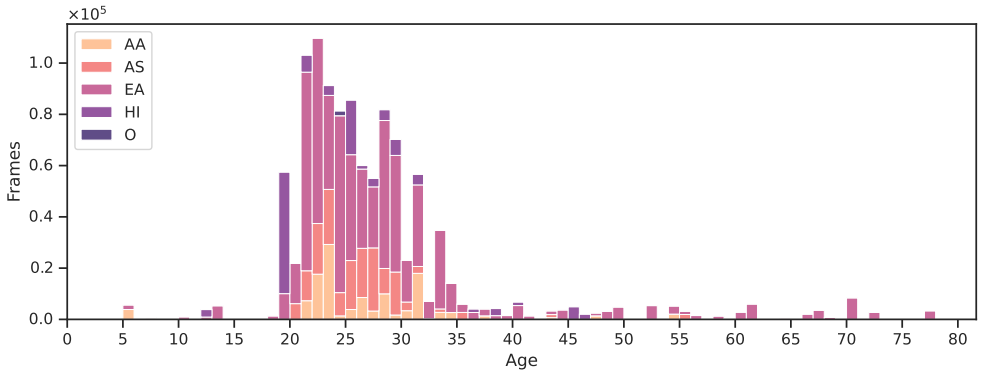

Figure 8: Age distribution of combined data, hue ethnicity (*AA*: African-American, *AS*: Asian, *EA*: Euro-American, *HI*: Hispanic, *O*: Other).