CYENS
CENTRE OF EXCELLENCE

Document Title	**Analysis of Algorithm Performance**
Project Title and acronym	DEtecting Stereotypes in human ComputAtioN Tasks (DESCANT)
Pillar	II. Sustainable RTDI System
Programme	Excellence Hubs
Grant Agreement	EXCELLENCE/0918/0086
Deliverable No.	5.2
Work package No.	5
Work package title	Demonstration System
Authors (Name and Partner Institution)	A. Lanitis (CYENS)
Contributors (Name and Partner Institution)	E. Christoforou (CYENS) A. Kafkalias (CYENS)
Reviewers	G. Demartini (UQ) J. Otterbacher (CYENS)
Status (D: draft; RD: revised draft; F: final)	F
File Name	DESCANT_Deliverable_5_2
Date	30/6/2022

ΙΔΡΥΜΑ ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

This project has received funding from Research and Innovation Foundation (RIF) under grant agreement EXCELLENCE/0918/0086.

Draft Versions - History of Document				
Version	**Date**	**Authors / contributors**	**e-mail address**	**Notes / changes**
1.1	20/4/2022	A.Lanitis	Andreas.lanitis@cut.ac.cy	First Draft
1.2	20/6/2022	A.Lanitis	Andreas.lanitis@cut.ac.cy	Revised Draft
1.3	30/6/2022	J. Otterbacher	jahna.otterbacher@ouc.ac.cy	Final Version

Abstract

This deliverable describes the development and evaluation of several image analysis algorithms. The algorithms were trained using the datasets described in deliverable D5.1. Specifically, face image classification experiments utilizing machine learning models trained using data annotated by different groups of annotators, are presented. The experiments aim to quantify the effect of annotation bias introduced by different groups of annotators, allowing in that way the understanding of the problems that arise due to annotation bias. The results of the experiments indicate that the performance of Machine Learning models in several face image interpretation tasks is correlated to the self-reported demographic characteristics of the annotators.

Keyword(s):	Algorithm development and evaluation, crowdsourced training data, face image analysis

Contents

1. Introduction

Over the last decade, the use of Machine Learning (ML) has increased dramatically [1] as numerous daily tasks are accomplished based on ML models. For example, ML has been used in recommendation systems, speech recognition, robot control, medical diagnosis, natural language processing, weather forecast, biometric authentication, text/image synthesis and for many other applications.

At its core, an ML model will only be as good as the data used for training the model. The main issue that relates to the quality of a training dataset is how well training samples represent the classes to be classified, in terms of quality and quantity. Furthermore, an important aspect of the training data is the quality of the annotation, as imperfections in the annotation process can influence the training data quality. Quite often, the annotation process requires human expertise, and as a result it is subjected to the expression of social stereotypes. This is because as social beings, humans are continuously engaged in a process of interpreting and forming impressions of others. However, cognitive heuristics often lead us to make trait inferences and evaluations of others that are based on very little concrete evidence (i.e., social stereotyping) [2]. For example, political candidates whose facial appearance is regarded as more accomplished, have a higher chance of winning the elections [3]. The process of data annotation can be influenced by social stereotyping and introduce bias in ML models, that eventually affects their performance.

In this study, we aim to quantify the effect of annotation bias, in terms of the performance of ML models, allowing in that way the understanding of the problems that arise due to the expression of social stereotypes in the annotation process. In particular, we compare the performance of ML models trained using data annotated by male annotators, female annotators, and annotators belonging to different (self-reported) racial groups. All groups of annotators annotated face images in relation to the several face image interpretation tasks including the tasks of gender recognition, race classification, attractiveness estimation, and trustworthiness estimation. The comparison of the performance of ML models trained using data annotated by different annotator groups, allows the derivation of conclusions related to the effects of social bias (stereotyping) in ML. To further emphasize the extent of the annotation bias problem, we also present results that show that it is possible to determine the group of annotators involved in the annotation process, by considering the annotation data provided.

Two main experiments were performed. The first set makes use of ML models trained using Python and Tensorflow, utilizing architectures primarily used for face image impetration tasks. Those models were tuned using transfer learning based on annotator specific data collected as part of this project. The second experiment was based on ML models trained using a publicly available software platform, the lobe.ai. In addition, a third experiment was run where a reverse process was adopted, in order to test whether it was feasible to predict annotator characteristics based on annotation data. For this experiment Python and

Tensorflow were used for training the ML models. For all experiments, the models were trained using data from the Chicago Face Database, annotated using were data collected using the ClickWorkers platform. More details about the Chicago Face Database and the annotation process are provided in deliverable D5.1

2. Experiment 1: Face Image Interpretation Using ML Models Trained Using Python and Tensorflow

We investigate whether the biases of the annotators can be found in the predictions of the ML models trained on labelled data. Specifically, we aim to explore whether these biases can be found in a variety of tasks including Attractiveness, Trustworthiness, Age, Gender and Race estimation from face images. We chose to use neural networks in our experiments because they have been shown to have exceptional results in a variety of disciplines and it is the most popular ML algorithm today. In addition to that, our literature review suggests that the state-of-the-art approaches on facial image datasets, use a combination and/or variation of Convolutional Neural Networks (CNNs). One problem we faced during training was that the dataset is very small (approximately 600 pictures). This is a problem because overfitting occurred. We used data augmentation to combat this problem. The data augmentation process included the tasks of:

1. Horizontal flip
   creates a mirrored image from the original one along the vertical direction.
2. Translation
   performs a random moving of the original image along the horizontal or vertical direction (or both). Padding zeros are added to the image sides
3. Random Rotation:
   applies a random rotation between -10° and $+10^{\circ}$ to the image
4. Image generation using deep GANs
   create new data with the same distribution of training data. The generator attempts to produce a realistic image to fool the discriminator, which tries to distinguish whether this image is from the training set or the generated set

Several deep network architectures, used for face image interpretation tasks were considered including: FaceNet [8], VGG-Face [9], DeepID[10], ArcFace [11], OpenFace [12] and DeepFace [13]. In all cases pretrained versions of the models were used, and these models were tuned using transfer learning based on the annotator specific data collected as part of the project.

Performance of pre-trained models

The following data were taken using a 70-30 split for the training-test set. The Data augmentation was only done on the training data. The training and testing of the classification tasks were done using stratified sampling. The results are shown below:

Model Name/ Task	Trustworthiness (MAPE*)	Attractiveness (MAPE*)	Age (MAPE*)	Gender (Accuracy)	Race (Accuracy)
FaceNet	9.61	17.57	12.44	0.94	0.85
VGG-Face	9.45	18.17	12.41	0.97	0.83
DeepID	9.32	17.17	14.02	0.91	0.84
ArcFace	9.82	19.31	12.84	0.97	0.82
OpenFace	9.69	19.22	12.71	0.86	0.82
DeepFace	9.84	20.90	14.92	0.92	0.85

*Mean absolute percentage error

As the table makes clear the pre-trained FaceNet [8] has the best general performance since it managed to be in the top 3 in all tasks. While a decent performance is achieved for the rest of the tasks. For this reason, all remaining experiments were run using the FaceNet architecture

Models trained using annotators from different age groups:

For this experiment we divided the annotators in two groups (Old and Young) according to their age. We set the threshold to be 40 years old. The results obtained for the different tasks are shown below:

Age estimation:

Metric/ Dataset	Baseline	Old	Young	Random
Mean absolute percentage error	10.85	27.20	60.98	88.18
Mean absolute error	3.01	9.51	25.21	43.95
Mean square error	15.41	144.35	857.17	2512.67
Mean distance from Baseline	0	5.37	10.74	21.40

Attractiveness estimation

Metric/ Dataset	Baseline	Old	Young	Random
Mean absolute percentage error	16.96	46.15	57.09	43.74
Mean absolute error	0.51	1.44	2.69	1.71
Mean square error	0.42	3.09	10.78	4.8

Mean distance from Baseline	0	0.73	2.07	2.98

Trustworthiness estimation:

Metric/ Dataset	Baseline	Old	Young	Random
Mean absolute percentage error	9.12	51.96	48.54	44.11
Mean absolute error	0.30	1.72	2.20	2.00
Mean square error	0.14	3.99	8.51	5.41
Mean distance from Baseline	0	0.68	2.31	0.92

Gender estimation

Metric/ Task	Baseline	Old	Young	Random
Categorical cross entropy	0.36	1.83	2.94	5.17
Accuracy	0.99	0.83	0.72	0.5

Race estimation

Metric/ Task	Baseline	Old	Young	Random
Categorical cross entropy	3.66	2.49	1.38	1.41
Accuracy	0.86	0.65	0.45	0.57

Models trained using annotators from different educational levels:
For this experiment we divided the annotators in groups according to their educational level. The results obtained for the different tasks are shown below:

Age estimation:

Metric/ Dataset	Baseline	Bachelor or equivalent	High-school graduate	Master or equivalent	Random
Mean absolute percentage error	10.85	36.11	83.93	86.11	64.22
Mean absolute error	3.01	16.16	25.05	33.30	24.94
Mean square error	15.41	447.85	739.36	739.36	999.31

Mean distance from Baseline	0	4.65	20.16	16.60	9.86

Attractiveness estimation

Metric/ Dataset	Baseline	Bachelor or equivalent	High-school graduate	Master or equivalent	Random
Mean absolute percentage error	16.96	48.83	46.39	33.87	44.25
Mean absolute error	0.51	2.18	1.2	1.15	1.00
Mean square error	0.42	8.54	2.14	2.64	5.43
Mean distance from Baseline	0	2.07	0.71	0.93	0.37

Trustworthiness estimation:

Metric/ Dataset	Baseline	Bachelor or equivalent	High-school graduate	Master or equivalent	Random
Mean absolute percentage error	9.12	62.46	29.47	26.45	43.75
Mean absolute error	0.30	3.15	0.99	1.14	1.71
Mean square error	0.14	13.75	1.56	2.4	4.82
Mean distance from Baseline	2.31	2.07	0.68	0.68	0.92

Gender estimation

Metric/ Task	Baseline	Bachelor or equivalent	High-school graduate	Master or equivalent	Random
Categorical cross entropy	0.36	2.32	1.67	3.65	3.44
Accuracy	0.99	0.75	8.9	0.83	0.43

Race estimation

Metric/ Task	Baseline	Bachelor or equivalent	High-school graduate	Master or equivalent	Random
Categorical cross entropy	3.66	2.62	1.41	1.41	10.16
Accuracy	0.86	0.63	0.52	0.54	0.33

Models trained using annotators from different genders:
For this experiment we divided the annotators in groups according to their gender. The results obtained for the different tasks are shown below:

Age estimation:

Metric/ Dataset	Baseline	Male	Female	Random
Mean absolute percentage error	10.85	37.40	6.3	76.67
Mean absolute error	3.01	10.21	1.74	23.35
Mean square error	15.41	114.00	3.75	880.79
Mean distance from Baseline	0	8.40	9.75	13.57

Attractiveness estimation

Metric/ Dataset	Baseline	Male	Female	Random
Mean absolute percentage error	16.96	28.85	20.00	44.12
Mean absolute error	0.51	2.01	1.99	2.00
Mean square error	0.42	8.14	1.00	5.41
Mean distance from Baseline	0	2.07	1.07	2.97

Trustworthiness estimation:

Metric/ Dataset	Baseline	Male	Female	Random
Mean absolute percentage error	9.12	22.58	5.13	43.74
Mean absolute error	0.30	1.03	0.2	1.71
Mean square error	0.14	1.75	0.004	4.81

Mean distance from Baseline	0	2.31	0.25	0.91

Gender estimation

Metric/ Task	Baseline	Male	Female	Random
Categorical cross entropy	0.36	0.88	1.95	13.39
Accuracy	0.99	0.97	0.90	0.46

Race estimation

Metric/ Task	Baseline	Male	Female	Random
Categorical cross entropy	3.66	1.54	1.72	1.41
Accuracy	0.86	0.79	0.78	0.25

Models trained using annotators from different races:
For this experiment we divided the annotators in groups according to their race. The results obtained for the different tasks are shown below:

Age estimation:

Metric/ Dataset	Baseline	Asian	Black	Latino	White
Mean absolute percentage error	10.85	98.57	97.43	98.05	98.05
Mean absolute error	3.01	35.48	19.48	25.98	33.48
Mean square error	15.41	1260.05	379.62	677.37	1185.42
Mean distance from Baseline	0	10.09	23.51	23.26	3.07

Attractiveness estimation

Metric/ Dataset	Baseline	Asian	Black	Latino	White
Mean absolute percentage error	16.96	89.35	89.72	61.42	86.56

Mean absolute error	0.51	4.48	4.48	0.98	3.48
Mean square error	0.42	21.14	20.12	1.22	13.17
Mean distance from Baseline	0	1.71	1.89	1.66	2.04

Trustworthiness estimation:

Metric/ Dataset	Baseline	Asian	Black	Latino	White
Mean absolute percentage error	9.12	66.60	89.93	88.55	70.43
Mean absolute error	0.30	2.99	4.98	3.99	2.99
Mean square error	0.14	12.98	27.11	16.17	15.20
Mean distance from Baseline	0	0.56	2.14	1.87	1.31

Gender estimation

Metric/ Task	Baseline	Asian	Black	Latino	White
Categorical cross entropy	0.36	0.96	1.09	4.98	1.12
Accuracy	0.99	0.91	1.0	0.67	0.33

Employment estimation

Metric/ Task	Baseline	Asian	Black	Latino	White	Random
Categorical cross entropy	3.66	4.9	1.52	1.41	1.41	
Accuracy	0.86	0.7	0.56	0.51	0.25	

The results from our study suggest that machine learning models indeed share the biases of the annotators whose labels they trained on.

This finding emphasizes the importance of deliberate selection of the people who annotate datasets and stratified sampling in crowdsourcing work. We believe that the results of this work

should be a communicated to the data selling industry as well as well as any business who aims to use data to make informed predictions in order to preserve the integrity of the company.

3. Experiment 2: Comparing the Performance of Annotator-Specific Classification Models

The aim of this experiment is to compare the performance of ML models trained using the collected data, as to quantify the extent of possible bias introduced by different groups of annotators.

3.1 Model Training

During the process of model training, nine different Deep Learning Models mod- els were trained for each of the four tasks of gender, race, attractiveness and trustworthiness classification. For each model trained, the training data used is the one provided by the six groups of annotators (Male, Female, Asian, Black, Latino, White). Furthermore, for each classification task a model was trained based on the ground truth provided with the Chicago Face Database, and an additional model was trained based on randomly annotated data. To compensate for the fact that the vast majority of the annotated samples were attributed to White annotators, a randomly selected subset of samples annotated by White annotators was selected, where the number of observations in that case was on par with the numbers of observations from Asian, Black and Latino annotators. The model trained using the subset of white annotators was called "Reduced White" (RWh).

Model training was done using the lobe.ai tool [6]. By using open-source Ma- chine Learning Architectures, the lobe.ai tool is able to automate Deep Learning classification tasks without the need to perform a rigorous manual model optimisation process, ensuring that all models under comparison are trained using exactly the same training and model optimisation procedures. Furthermore, the lobe.ai tool is able to achieve an excellent performance at low computational costs. However, although lobe.ai allows the export of trained models for use in conjunction with the most popular deep learning libraries, it does not provide explicit details of the model architecture and/or the training algorithms used for training the models. After loading the data with the appropriate labels, the lobe.ai tool needs approximately around 10-15 minutes in training and optimizing the models when the model training procedure was run on an AMD Ryzen 3600 6-Core Processor with 16 GB RAM.

3.2 Results and Discussion

Details of all models trained in terms of the number of samples and the performance achieved on the train and test data is shown in table 1.

Table 1: Models’ train and test accuracy for each classification tasks divided by the respective annotation categories.

		Gender		Race		Attractiveness		Trustworthiness	
Model	Num of Samples	Train	Test	Train	Test	Train	Test	Train	Test
Ground Truth(GT)	597	99	97	93	91	77	76	62	60
All Annotators(AA)	597	99	98	86	80	62	53	56	36
Male(Ma)	596	92	94	72	32	66	53	59	39
Female(Fe)	592	93	95	84	76	68	50	56	38
Asian (As)	106	94	88	90	65	36	28	44	27
Black (Bl)	165	95	81	80	56	74	36	75	36
Latino (La)	114	89	86	75	76	49	23	70	34
White (Wh)	597	94	93	82	69	70	47	57	39
Reduced White (RWh)	128	92	88	79	65	79	45	62	31
Random (Ra)	597	67	37	54	17	47	27	62	36

As expected models built using the ground truth data outperform the rest of the models. On the other hand, in most cases models built on random data have the worst performance. Apart from the race classification task, models trained using data annotated by male and female annotators have similar performance indicating that, for the tasks considered, the annotation process by male and female annotators leads to models with similar performance. However, models trained using data annotated by annotators belonging to different racial groups display increased diversity in performance.

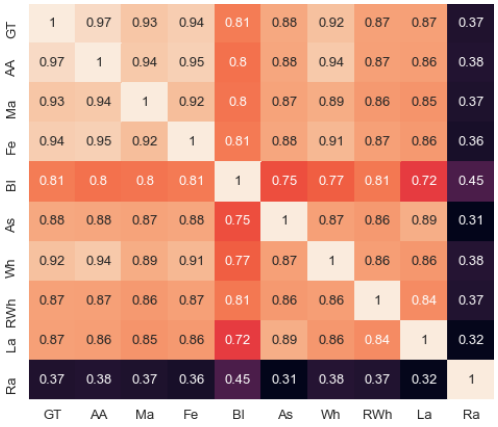
With the introduction of deep learning and convolutional networks, tasks such as gender classification are now considered trivial for ML problems with expected accuracy of around 95%. However, models trained using data annotated by annotators from different racial backgrounds resulted in worse performance compared to the models built using data from annotators of different genders. Among the classification problems considered, the task of trustworthiness estimation received the lowest classification rates, implies that trust cannot be easily determined based only on facial appearance. For the attractiveness task, the models trained based on the Asian and Latino annotators achieve the worst performance, on par with the performance achieved by the models based on random annotations. This observation can be linked to different attitudes related to attractiveness cultivated in different cultures [7].

Comparing the results of the models built from the entire dataset of the White annotators (’White’ model) against the models built with a randomly selected subset of images annotated by White annotators (’Reduced White’ model), it is observed that although the models trained using reduced samples achieved lower performance, no major differences in relation to the comparison against models trained using Asian, Latino and Black narrators is observed. Therefore, we can conclude that the results related to the performance of the model trained using White annotators, are not attributed to the higher number of samples used during the training process.

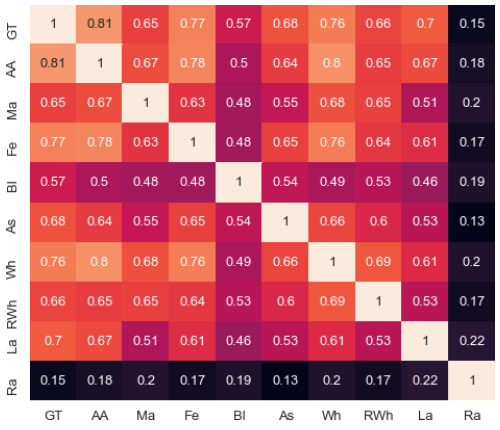
The correlation matrices shown in figure 2 demonstrate the percentage agreement between the classification performance achieved by different models, for each task considered in the evaluation. For the tasks of gender and race classification (figure 1.(a) and figure 1.(b)) most of the models achieve high percentage agreement between each other, with an average of around 90% and 75% respectively. Excluding the models based on Black annotators and random models, the rest of the models have a similar agreement. Even though models based on Black annotators have a high percentage of test accuracy, it is clear that they disagree the most with the rest of the models with an average of 80% for gender and 50% for race. The contradiction between the agreement of models is attributed to label bias introduced during the annotation process. Based on

data from figure 1, it is evident that Black annotators classify race and gender in a different way and this impacts the result of the respective models.

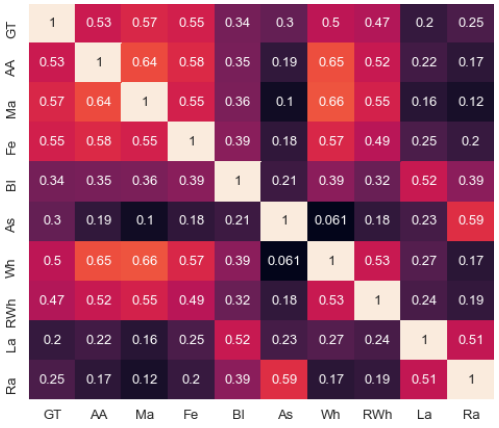
All models built for the trustworthiness estimation task, have very low agreement with each other. Furthermore, classification results for the task of trustworthiness estimation display high diversity among different groups of annotators, indicating the perception of those attributes varies from gender to gender and race to race. The results clearly demonstrate that trustworthiness estimation is subjected to bias due to the annotation process, hence a proper annotation procedure that involves annotators from different groups need to be employed to produce training data suitable for this challenging task. Except for the task of trustworthiness estimation, the models trained based on all white annotators and the models trained using the reduced subset of white annotators have high level of agreement. This indicates that when compared with the annotation bias introduced by annotators from different backgrounds, the bias introduced by different sizes of training samples is less important for the tasks of gender, race and attractiveness classification.



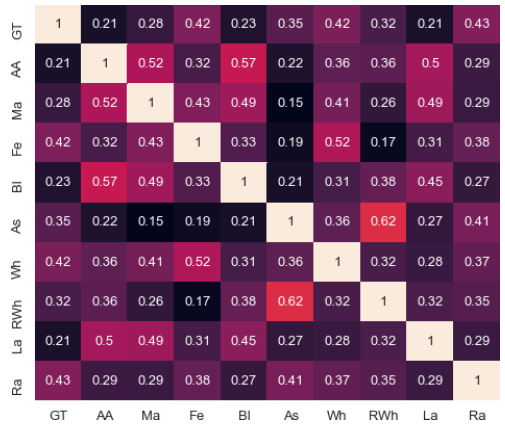
(a) Gender



(b) Race



(c) Attractiveness



(d) Trustworthiness

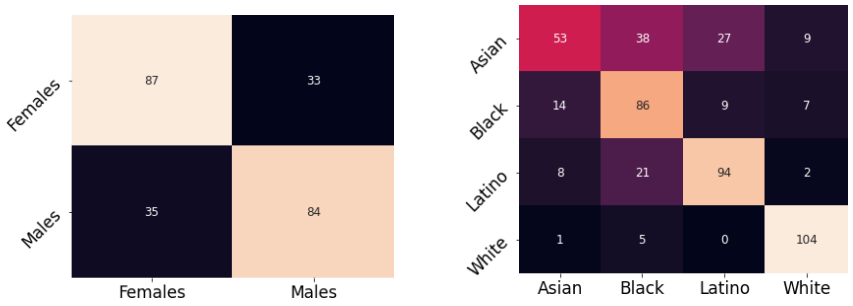
Fig. 1: Correlation Matrices between the models of each category. Light colour indicates high percentage agreement while darker colours low percentage agreement.

4. Experiment 3: Predicting annotator groups based on annotations

The results of experiment 1 show clear differences in performance, and disagreement between models trained using data annotated by different groups of an- notators. To further investigate this phenomenon, the possibility of predicting the gender and the race of an annotator, based on their respective annotations was examined. In this context two classification models were trained. Each of those models take as input the ground truth values of gender, race, attractive- ness and trustworthiness for each sample, along with the annotation provided by each annotator, using the data collected through the clickworkers platform (see Deliverable 5.1). The models trained use a Multilayer Percepton (MLP) architecture with eight inputs, four fully connected layers of 128 neurons with relu activation, a fully connected layer with 64 neurons and relu activation, and an output fully connected layer with a sigmoid activation. The outputs of the model correspond to the gender and race of an annotator.

Once trained, the models are able to identify correctly the gender with a test accuracy of 70% for the race and 66% for the gender. The confusion matrices below in figure 3 demonstrate the results for the prediction of the annotators.

A cross validation with K folds, where K=30, was run, in combination with a hypothesis z-test, to examine if the trained models can be used to predict the attributes of annotators with better accuracy than random guesses. The results indicate that there is a statistically significant improvement between the predicted labels of the trained models when compared to random guesses, for both the gender (z=5.57 | a =0.01) and race (z=20.54 | a =0.01). Based on these results, it is evident that the bias in the annotations is reflected as bias in the models predictions, and using these predictions it is possible to reverse engineer the process and identify attributes of an annotator.



(a) Gender (b) Race

Fig. 3: Confusion matrices for both models created for predicting the features of the annotators. Light colour indicates a high prediction count while darker colours a lower prediction count.

5. Conclusions

Experimental results presented in this deliverable demonstrate that the perception of characteristics such as attractiveness and trustworthiness vary as a function of annotator demographics (gender and race). In fact, sometimes machine learning models trained to classify those attributes have higher levels of agreements with models based on random data instead of another category of annotators. Even bi- nary gender classification, which can be considered trivial in face interpretation, show different results when models are trained on data from annotators from different racial backgrounds. Furthermore, the bias in the models predictions makes it possible to reverse engineer and identify attributes of an annotator.

Based on the results obtained, it is evident that computer vision tasks that rely on training data annotated by humans could be heavily influenced by social stereotyping, that can cause biased performance. The work presented in this study provides quantitative results indicating the extend of the problem in several classification tasks, against the groups of annotators used, providing in that way useful insight for researchers involved in similar classification tasks.

The results of this study are demonstrated through an interactive tool at http://recant.cyens.org.cy/, so that the results of this project can be used by Machine Learning practitioners and students, as training material to anticipate the dangers of annotation bias.

6. Bibliography

1. Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science 349(6245), 255–260 (2015)
2. Jussim, L., Nelson, T.E., Manis, M., Soffin, S.: Prejudice, stereotypes, and labeling effects: Sources of bias in person perception. Journal of personality and Social Psychology 68(2), 228 (1995)
3. Said, C.P., Sebe, N., Todorov, A.: Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. Emotion 9(2), 260 (2009)
4. Chicago Face Database. https://www.chicagofaces.org/, accessed: 22-02-2022
5. Clickworker crowdsourcing. https://www.clickworker.com/, accessed: 22-02-2022
6. Lobe.ai webpage. https://www.lobe.ai/, accessed: 22-02-2022
7. Rhodes, G., Lee, K., Palermo, R., Weiss, M., Yoshikawa, S., Clissa, P., Williams, T., Peters, M., Winkler, C., Jeffery, L.: Attractiveness of own-race, other-race, and mixed-race faces. Perception 34(3), 319–340 (2005)
8. [F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.
9. Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 67-74, doi: 10.1109/FG.2018.00020.
10. Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes[C]//Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014: 1891-1898.
11. J. Deng, J. Guo, J. Yang, N. Xue, I. Cotsia and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-1, 2021. Available: 10.1109/tpami.2021.3087709.
12. T. Baltrušaitis, P. Robinson and L. Morency, "OpenFace: An open source facial behavior analysis toolkit," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1-10, doi: 10.1109/WACV.2016.7477553.
13. Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.