Overcome Ethnic Discrimination with Unbiased Machine Learning for Facial Data Sets*

Michael Danner[1,†][a], Bakir Hadžić[2,†], Robert Radloff[2], Xueping Su[3][b], Leping Peng[4],
Thomas Weber[2] and Matthias Rätsch[2][c]

[1]*CVSSP, University of Surrey, Guildford, U.K.*

[2]*ViSiR, Reutlingen University, Germany*

[3]*School of Electronics and Information, Xi'an Polytechnic University, China*

[4]*Hunan University of Science and Technology, China*

Keywords: Unbiased Machine Learning, Fairness, Trustworthy AI, Acceptance Research, Debiasing Training Data, Facial Data Sets, AI-Acceptance Analysis.

Abstract: AI-based prediction and recommender systems are widely used in various industry sectors. However, general acceptance of AI-enabled systems is still widely uninvestigated. Therefore, firstly we conducted a survey with 559 respondents. Findings suggested that AI-enabled systems should be fair, transparent, consider personality traits and perform tasks efficiently. Secondly, we developed a system for the Facial Beauty Prediction (FBP) benchmark that automatically evaluates facial attractiveness. As our previous experiments have proven, these results are usually highly correlated with human ratings. Consequently they also reflect human bias in annotations. An upcoming challenge for scientists is to provide training data and AI algorithms that can withstand distorted information. In this work, we introduce AntiDiscriminationNet (ADN), a superior attractiveness prediction network. We propose a new method to generate an unbiased convolutional neural network (CNN) to improve the fairness of machine learning in facial dataset. To train unbiased networks we generate synthetic images and weight training data for anti-discrimination assessments towards different ethnicities. Additionally, we introduce an approach with entropy penalty terms to reduce the bias of our CNN. Our research provides insights in how to train and build fair machine learning models for facial image analysis by minimising implicit biases. Our AntiDiscriminationNet finally outperforms all competitors in the FBP benchmark by achieving a Pearson correlation coefficient of $PCC = 0.9601$.

1 INTRODUCTION

In recent years, the use of artificial intelligence has proven to solve a wide spectrum of technical problems. Especially in the high-tech sector and in knowledge intensive industries, machines and intelligent algorithms turned from clunky tools to sophisticated systems performing various complex tasks today (Arslan et al., 2021). In today's global war of talents, companies are hunting for the best employees with specific requirements of skills and personal traits to achieve competitive advantage in their field (Grant, 1991). In this context, a wide range of research has been conducted to understand the evolutionary basis of beauty and determine the bias of attractiveness in the job hiring process (Little et al., 2011; Chiang and Saw, 2018). Companies desire an efficient and objective recruitment process with the preferred outcome of finding the best job candidates and stay compliant with regulations and ethical aspects. Artificial intelligence has the potential to support these goals by minimising the risk of bias in decision making in order to be a relevant and trustworthy partner for humans in the future.

1.1 Motivation

In 2016 *Beauty.AI*, a Hong-Kong based technology company, hosted the first international beauty contest judged by artificial intelligence (beauty.ai, 2016) but the results were heavily biased, for example, against dark-skinned subjects (Levin, 2016). "Machine learning models are prone to biased decisions, due to bi-

[a] https://orcid.org/0000-0002-8652-6905

[b] https://orcid.org/0000-0003-1306-8453

[c] https://orcid.org/0000-0002-8254-8293

[†] Both authors contributed equally

[*] This work is partially supported by a grant of the BMWi ZIM-FuE programs, no. KK5007201LB0.

Danner, M., Hadžić, B., Radloff, R., Su, X., Peng, L., Weber, T. and Rätsch, M.
Overcome Ethnic Discrimination with Unbiased Machine Learning for Facial Data Sets.
DOI: 10.5220/0011624900003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 464-471
ISBN: 978-989-758-634-7; ISSN: 2184-4321

ases in data-sets" (Sharma et al., 2020). Biased training data potentially leads to discriminatory models, as the data sets are created by humans or derived from human activities in the past, for example hiring algorithms (Bogen, 2019). The purpose of Facial Beauty Prediction (FBP) research is to classify images mimicking subjective human judgements. Investigations related to machine perception in a ground-truth free setting show that the data source depends on the measurement of human perception (Prijatelj et al., 2020). Therefore, artificial networks need a process to determine labels of the average person's judgement. Our data analysis has already proven that people consider their own ethnicity to be more attractive than others (Gerlach et al., 2020). With this tendency, it becomes difficult to generate input data to train a machine learning algorithm, which assesses attractiveness without bias. This is a highly relevant topic in machine learning, which has a technical component to solve and trigger ethical debates about discrimination. To tackle this issue, we used our recently published balanced training data set (Gerlach et al., 2020). with minimal bias between European and Asian aesthetic predictions from a convolutional neural network (CNN). Based on this training data, we created a model capable to achieve an equally distributed performance over all classes for those different ethnicities. For the first time, we could prove, that this resampled and balanced training data leads to a debiased AI model for a fair facial aesthetic prediction among different ethnicities. The main contribution of this work is the novel approach towards an unbiased machine learning among different ethnicities to build a fair and trustworthy AI model, by applying a mixed data set, which consists of real images together with synthetic data.

1.2 Acceptance of AI

AI-based technologies are promising tools to optimise the process of recruiting, assisting recruiters in their routine work and thus increasing the efficiency of the whole recruiting process(Ahmed, 2018; Reilly, 2018). However, the extent of willingness to accept and use AI-enabled recruiting among the actual applicants is widely uninvestigated until now (Laurim et al., 2021). Consequently, one of the goals of the present study was to examine acceptance factors of AI use in recruiting process among university students. A total of 559 participants (49% female) completed the online survey. The results showed that AI acceptance is dependent on the following: a contact person should be available (91%), process should be transparent (86%) and ensure data protection (83%). Even though the minority (35%) of participants is generally worried about AI utilisation, however, only 52 % stated that they would support AI-enabled recruiting procedures (Schlick and Reich, 2021). Our survey demonstrated that the acceptance of AI-enabled systems heavily depends on the features of the provided AI. The AI system should make fair decisions that are transparent and consider individual personality traits while keeping their data safe and protected. Potential sources of worries and doubts regarding the AI among the applicants must be adequately addressed in future studies. First prerequisite for usage of AI systems is that they are accepted by the users. Therefore, one of the central goals of the following study is to design a fair and unbiased AI.

2 STATE OF THE ART

While research on the estimation of images or portraits is not a new trend, it has gained increasing attention since the emergence of artificial intelligence (Zhang and Kreiman, 2021). Although, AI is undoubtedly the best solution for many applications like autonomous driving or image classification, applications that are affected by unconscious bias, like beauty prediction (Dornaika et al., 2020), tend to reflect a bias that is likely to be prevalent within given data sets. Especially, when people's subjective preferences play a role, such as in attractiveness judgement (Shank and DeSanti, 2018) or human resource evaluation (Lloyd, 2018), bias is almost certain to happen. (Carrera, 2020) conducted research on the implication of racism in image databases that analysed the association of aggressiveness, kindness, beauty and ugliness with different images and found that the decisions of many people are affected by subconscious racism. Since researchers are aware of such effects, they found different ways to reduce subconscious bias in machine learning. If the problem originates from the given databases - either the databases or the training needs to be changed.

3 BIASED AI

3.1 Bias from Human Indications

Our latest data set included 12,034 images of people from different social and ethnic backgrounds that we collected during the period of eight years long period working on this topic. One part of the data was collected from the students of our partner university

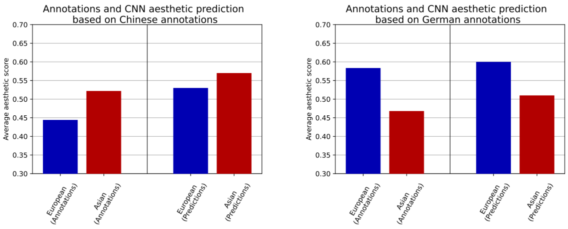

Figure 1: AntiDiscriminationNet is trained on annotations from students at a German and at a Chinese university. The trained network follows the bias from the annotations.

in China. This number of images provided 5.4 million annotations where different annotators evaluated the attractiveness of persons in the presented picture. Our aim in that study (Gerlach et al., 2020) was to empirically test if data manifest any implicit bias regarding ethnicity or some other relevant characteristics of evaluated pictures. Based on the literature research that we conducted, we hypothesised that the evaluated attractiveness of facial pictures in the Asia-Europe data set by annotators in China and Germany is implicitly biased. More precisely, we assumed that annotators are evaluating more attractive images of persons that are having the same ethnic background as them. To test this hypothesis, we separated results from annotators, divided them based on their racial background, and then compared their results. Germans represented European annotators, while Chinese represented Asian annotators. The results of the conducted analysis indicated that our hypothesis was supported. As we can see in Figure 1. European annotators evaluated European faces as more attractive, while Asian evaluated Asian faces as more attractive.

3.2 Artificial Intelligence as Facial Aesthetics Predictor

Current state-of-the-art results of Facial Aesthetics Predictor system are presented in this subchapter. Afterwards, we introduced our AntiDiscriminationNet Predictor. For the prediction of facial aesthetics, we used convolutional neural networks (CNN).

Related Work. With the introduction of CNNs and large-scale image repositories, facial image and video tasks get more powerful (Krizhevsky et al., 2017; Zeiler and Fergus, 2013; Deng et al., 2009). Xie et al. (Xie et al., 2015) present the SCUT-FBP500 dataset, containing 5500 subjects with attractiveness ratings. Since ”FBP is a multi-paradigm computation problem” the successor SCUT-FBP5500 (Liang et al., 2018) is introduced in 2018, including an increased database of 5500 frontal faces with multiple attributes: male/female, Asian/Caucasian, age and beauty score. Liang et al. (2018) have evaluated their database ”using different combinations of feature and predictor, and various deep learning methods” on AlexNet (Krizhevsky et al., 2017), ResNet-18 (He et al., 2015) and ResNeXt-50 and achieved the Pearson correlation coefficient $PCC = 0.8777$; with p value being statistically significant at $p < 0.01$, mean average error $MAE = 0.2518$; root-mean-square error $RMSE = 0.3325$ as a benchmark. In summary, it can be said that all deep CNN models are superior to the shallow predictor with the hand-crafted geometric feature or appearance feature (Liang et al., 2018).

Benchmark Data Set. The SCUT-FBP 5500 data set is a small data set for deep learning tasks. Therefore, it is an even greater challenge to train soft features like aesthetic or beauty. In order to measure the accuracy of the network and to be comparable to recent experiments in facial beauty prediction, we calculated the Pearson correlation coefficient, mean absolute error (MAE) and root mean square error (RMSE).

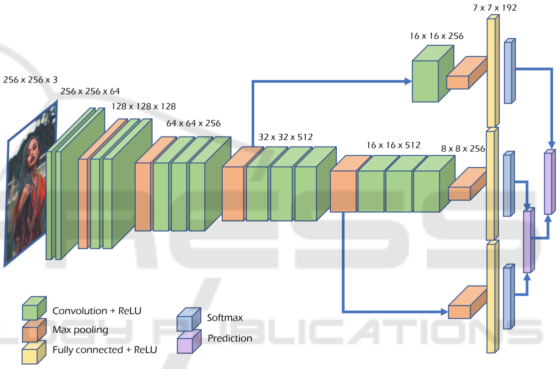

Figure 2: The architecture of AntiDiscriminationNet is based on the VGG Face architecture and is expanded by two separate skip connections. At the end, the predictions of the differently convoluted feature vectors are merged.

AntiDiscriminationNet Predictor Architecture. The VGG Face architecture (Simonyan and Zisserman, 2015) was the basis of our AntiDiscriminationNet. Inspired by an idea of the paper from (Shelhamer et al., 2017) we then added modifications to the network by exploiting feature maps from the third and fourth convolution blocks. Since the size of the feature maps differed from the size of the resulting feature vector, we implemented an additional max pooling layer to achieve the wanted output. For the predictions of the network, we concatenated the softmax results into a single feature vector as shown in Figure 2. Our proposed network achieved the Pearson correlation coefficient $PCC = 0.9601$; with p-value being statistically significant at $p < 0.01$, which indicated an almost linear correspondence between annotations and predictions. Our training results have very high accuracy and outperform state-of-the-art results. More detailed results, and comparisons with other networks are presented in our previous work (Danner et al., 2021). The normalised mean square error was $nMSE = 3.896$ and the normalised root mean square error was $nRMSE = 5.580$. These are measurements of the average error of the predicted labels, which were used to evaluate the accuracy of the network. The results are normalised because there are different data sets with different score ranges.

Re-annotation of SCUT-FBP5500 Data Set. Since 2013, for our study of facial aesthetics, we conducted online surveys on multiple image data sets where thousands of students and their relatives participated (Gerlach et al., 2020). With this process we have been able to gather enough data to train a convolutional neural network with the goal to improve facial beauty prediction. During training convolutional neural networks (CNN) on this data, we recognised a large bias in this data. This led us to evaluate the annotations from Chinese and German universities and take a closer look at the bias. We conducted statistical analysis to see if the trained network would also produce bias during aesthetic prediction. Results of mentioned former paper (Gerlach et al., 2020) indicated a statistically significant difference between CNN aesthetic prediction based on Chinese and German annotations. Results are presented in Figure 1. When based on Chinese annotations, CNN predicted a higher aesthetic score for Chinese annotations. Based on German annotations, CNN predicted a higher aesthetic score for German annotations. These results empirically revealed that the trained network reflected the same bias as human annotators. After this revelation, our next step was to train unbiased AI.

4 TRAINING OF UNBIASED AI

In general, there are three main paths to reach the goal of unbiased predictions: fair pre-processing, fair in-processing, and fair post-processing (Bellamy and et. al., 2018). Within this paper, we present two approaches based on those paths to train an unbiased network with biased data for FBP. The first approach relies on data pre-processing before training to introduce fairness - we call it ”balanced training”. The second approach relies on a categorical cross entropy loss function, for the network to learn the bias and decrease it. Those processes are explained in the following sections.

4.1 Data Set and GAN-Images

We analysed the data that we gathered with our Analysis Toolbox and could measure a significant bias within the prediction of aesthetics through different ethnicities. Therefore, training a network with the goal to create unbiased results is still a challenge in deep learning tasks. In the following we will first describe our data set blend and the accompanying Analysis Toolbox and we explain how we used a GAN to create artificial portraits with European and Asian ethnicities.

Starting in 2017, we used the Asian-European data set SCUT-FBP (Xie et al., 2015; Liang et al., 2018) to evaluate biased annotations from Chinese and German universities. Since the SCUT-FBP 5500 dataset is a small dataset for deep learning tasks, we used data augmentation methods to enlarge the sample size of the training set by generating GAN images with either Asian or European or mixed images as input and new synthesised images as output. This augmentation method proves superior to geometric transformations like cropping and rotating. All images are pre-processed, by normalisation methods to harmonise face pose, facial landmark positions and image size.

For the purpose of a detailed analysis, we blended multiple data sets in the domain of facial aesthetics together. In total, this data set included 12,034 portrait images from persons of different ethnicities with individual social backgrounds. These images are labelled and annotated in surveys over a period of eight years with a total number of 5.4 million annotations. Beside that, recently we added the FairFace (Kärkkäinen and Joo, 2019) database, which includes male and female portraits of seven different ethnic groups. The synthesised Eurasian images are artificially generated with StarGAN v2 (Choi et al., 2020) to determine the influence of the biased view of annotators on aesthetics of persons from different ethnicities.

After annotating the data set, the unconscious bias in the annotations can be uncovered. Figure 3 shows the biased average score of our networks on the SCUT-FBP data set and the Eurasian data set. Figure 4 illustrates the analysis on the distribution of aesthetic score and age for Asians, Europeans and three mixed-racial subgroups. The different group annotation points are displayed in different colours. We calculated the following metrics for each group cluster i: Horizontal dashed lines are average attractiveness values $\overline{a}_i$. Vertical dashed lines are average age values $\overline{y}_i$. As can be seen, the interval of $\overline{a}_i$ has a small span, however the interval of $\overline{y}_i$ has a significantly larger span. Each $\overline{a}_i$ and $\overline{y}_i$ values intersection point forms

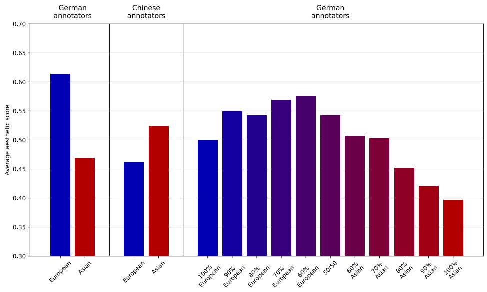

Figure 3: Unconscious bias towards ethnic aesthetic of either German or Chinese annotators. Left: average aesthetic score on SCUT-FBP by German annotators, middle: average aesthetic score labelled by Chinese students, right: aesthetic scores on the Eurasian data set annotated by German students.

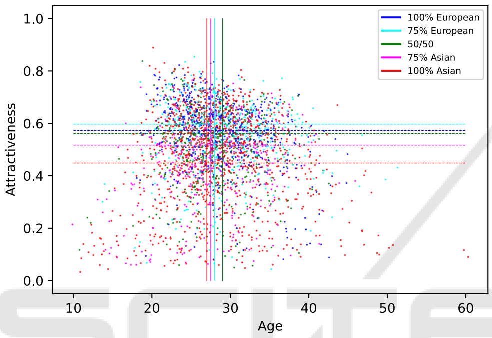

Figure 4: Biased correlation between attractiveness, age and ethnicity by German annotators. In an ethical, fair network the attractiveness for equal age groups would be the same. This would be represented in the figure by the same height of the lines for equal age groups.

an per group attractiveness-age-factor $AAF_i = \overline{a}_i / \overline{y}_i$. In a fair machine, these AAF_i points would be closer together, as the $\overline{y}_i$ span is small. This idea is further elaborated in Section 4.

4.2 Training and Data Pre-Processing

In our first approach of training the network we applied pre-processing and resampling to the input data, which is explained in the following paragraphs.

This paper proposes a way to create a fair network with this biased data. Therefore, the bias must be identified in the ground truth labels of the data set and divided into two subsets. The first subset (German annotations) confirms and increases the existing bias whereas the second subset (Chinese annotations) consists of the contrary prejudices (annotation bias). Afterwards, a GAN then generates synthetic images, which are a gradation of the mixture of the first and second subset. This selection bias leads to the best balanced results of the generated images. This knowledge can then be applied back to the original data set.

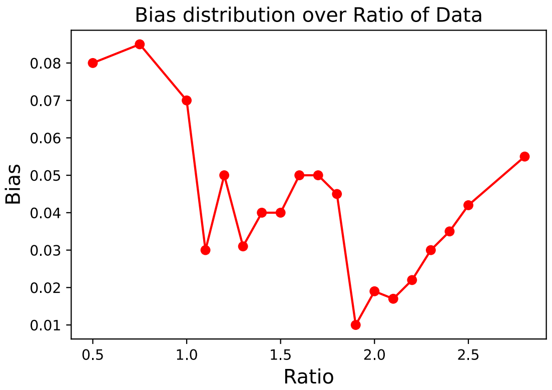

Figure 5: Correlation of the bias over the ratio of German and Chinese annotations. The least bias here is at the ratio of 1.9.

In our training process, we have a clear bias in the annotations, as shown in Figure 1 and measured in the analysis of the data. If we train our network based only on these labels, it follows the data and replicates the bias from the annotations. In the next training, we added the annotations from the Chinese and German annotators and trained the network on an equal distribution of those annotations (Ratio: 1.0). The average aesthetic rating of European and Asian faces is still biased, however not as strong as in the previous experiment.

In this experiment, balancing the training data meant finding the minimum by concatenating the German annotated subset g with the weighted ω Chinese annotated subset c. The goal in this approach was to level the average aesthetic scores $\overline{g}$ and $\overline{c}$ for the generated predictions g_i and c_i. The network bias B is then defined by

$$B = \frac{1}{2n+1} \sum_{i=0}^{n} |\overline{g} - g_i| + \omega |\overline{c} - c_i|. \qquad (1)$$

Starting from a ratio of 1:1, in which German and Chinese annotations are distributed equally, we gradually increased the weight of the Chinese annotations. In our experiment we varied the ratio from 2:1 to 1:3.2 for German annotations to Chinese annotations. Each training step and the corresponding bias over the ratio is shown in Figure 5. Determining the minimum in Figure 5 is equal to finding the least biased network. It is visible that a ratio of 1:1.9 produced the least biased network for this experiment and its results are shown in Figure 6. This means the Chinese annotations are weighted nearly double the amount than the European annotations. Limitations of this approach are that information about the structure of the underlying latent features are unknown and balancing the network requires a lot of time and work. Therefore,

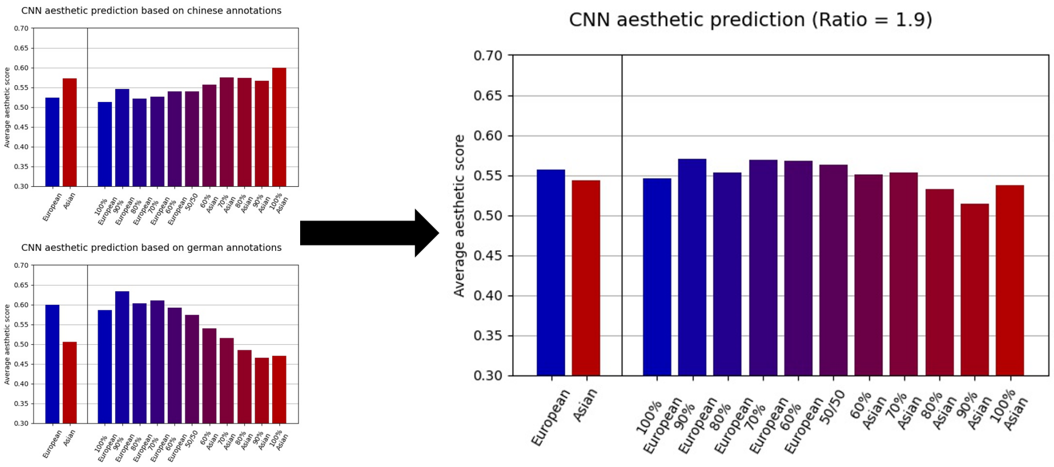

Figure 6: CNN aesthetic prediction with equalised distribution of training data. The charts on the left side show the prediction of the network if it is only trained on Chinese or German annotations. On the right side, the prediction of the network, which was trained on the biased data is shown. All bars have more or less the same height and only differ minimally. This means, that we could eliminate most of the bias in the training data, by balancing and we can assume that this trained network is fair.

we additionally propose another approach, described in the following section.

4.3 Debiasing Neural Network

4.3.1 Training Network Features

To achieve the first results on unbiased aesthetic estimation, we used the existing VGG-Face framework in Keras with TensorFlow and adjusted it. The network consists of 11 blocks, each containing a linear operator and followed by one or more non-linearities such as ReLU and max pooling (Parkhi et al., 2015). We applied transfer learning here and used the pretrained model for Face Recognition (Parkhi et al., 2015). Building up on the face recognition, attractiveness estimation is similar to age estimation (Gyawali et al., 2020) performed by observing the facial features from portraits. The convolutional layers in the network are followed by a rectification layer (ReLu) as in (Krizhevsky et al., 2017). We used the Adam optimiser (Kingma and Ba, 2017). The input to our network is a face image of the size $256 \times 256 \times 3$, and it uses Zero-Padding around the edges, to ensure that the image information on the edge is not lost. Our input data is split into 60% train and 40% test data. The convolutional layers parameters of VGG-Face are not changed and kept frozen during the training. We used a dropout of 50%, and as it is a regression problem our final layer had to be the size of 1. To classify the aesthetic score, a softmax activation function is used in the final layer. As a loss metric, we used the mean squared error and to compare our networks we also calculated the Pearson correlation and the root mean squared error.

4.3.2 Balanced Training

The process and the effect of the ratio on the average aesthetic rating is shown in Figure 5. By modifying the ratio of the annotations, a minimum is determined that illustrates the lowest difference between the average aesthetic prediction of Asian and European faces. This represents a specific loss function for our network that maps bias onto measurable values. To remove bias from our network, we calculated the difference between European and Asian aesthetic predictions and found the global minimum. The minimum of the average aesthetic score between Asian and European faces is located at a ratio of 1:1.9 where the average aesthetic score differs by about 5%. We created a model with a fair performance over all classes of different ethnicities as shown in Figure 6. This proves, that by re-sampling and balancing the training data a less biased AI can be created. This process created a less-biased AI in FBP tasks.

Our results are displayed in Figure 6 where all bar charts have a similar height and the FBP score is considerably less biased. Not all bars have the exact same height, this is due to some background noise. Real world data usually contains noise which affects tasks such as classification in machine learning (Gupta and Gupta, 2019). This noise also affects our aesthetic prediction. However, with those minor differences, we can consider our network as unbiased and therefore fair. As we used a factor-based approach to multiply the annotation data, this noise is present over all ratios. Only the difference of the averages increases or decreases within the variations of the ratio.

4.3.3 Removing Bias Using Clustered Labels

A more sophisticated approach in getting rid of the bias in training data is our second approach. Within this approach we are developing a new method to reduce the bias in the training data. This method consists of a deep learning network that is trained on the original learning task within the data set, and then minimises the bias inside the learned latent distributions using a specially adapted loss function.

Each data record contains a list of labels $a = a_1,...,a_n$, which are to be debiased, and a further list of labels $b = b_1,...,b_n$. In this example we remove the bias from the ethnical label a_1 and preserve the age, profession, hair colour and skin complexion labels. The network evaluates all attributes of the data set during the training and groups all objects according to the attributes b in clusters.

Within each subgroup the difference between the ethnical mean value $\overline{a_1}$ represents the bias. A non-linear operation, similar to the gamma correction in image systems, is then applied to the ethnic label to preserve the range of the values and bring the differences closer together. These differences for all clusters are the measure of the loss function, which is implemented as categorical cross entropy loss and should be minimised during training. With this we present a universally adaptable method to make any network fairer according to given labels.

5 FURTHER ETHICAL CONSIDERATIONS

Societal benefits can arise from aesthetic prediction when trustworthy AI models are used. We presented a method to eliminate bias in facial attractiveness prediction and this method can be transferred to other similar networks and use cases. For example, in the future, a possible implication of our unbiased AntiDiscriminationNet is supporting the recruitment processes or plastic aesthetic surgeries in the medical domain. Applying those AI models needs to be discussed in the light of a benefit and risk assessment. The implementation of machine learning models in various future tasks must be accompanied by effective measures, long-term ethical considerations and transparency. Training of an unbiased model on biased data will be a constant challenge in machine learning, especially in the field of aesthetic judgement and other fields where underrepresented ethnic groups are common. Pragmatical regulations and an open-source mindset will reduce the implication that algorithms may become a major threat of discrimination on a level of gender, sex, and ethnicity. Applying a debiased prediction model like AntiDiscriminationNet is a starting point for future research and continuous ethical evaluation. The implication and future research in this domain are driven by the adoption of AI models which allow us to screen facial images in a high throughput manner when needed and within the regulations. Beauty and aesthetic attractiveness predictions raise ethical questions and concerns. This debate needs to be held on an ongoing and open basis, ideally with a diverse group of stakeholders. Ultimately, the well-known phrase ‘beauty is in the eye of the beholder’ will stay as a universal rule for machines and humans in attractiveness judgement.

6 CONCLUSION

In this research we used experimental methods to systematically demonstrate how human implicit bias affects the decision-making of artificial intelligence and found a way to eliminate the implicit bias of artificial intelligence. Additionally, we improved the fairness of the prediction towards an equally distributed prediction between different ethnicities. Moreover, the Pearson correlation coefficient of $PCC = 0.9601$, which denotes a nearly linear correspondence between annotations and predictions, was reached by our trained network. Our training results are more accurate than those obtained in recent studies in this area. In summary, two main contributions of this paper are AntiDiscriminationNet for facial image analysis and a new approach towards bias-free machine learning models. Bias-free decision making is a challenging problem in machine learning tasks, yet it yields the great potential to be one of the most significant strengths of AI. Future work on this topic should focus on scaling our approach on larger, more diverse data sets and in other use cases.

REFERENCES

Ahmed, O. (2018). Artificial intelligence in hr. *International Journal of Research and Analytical Reviews*, 5(4):971–978.

Arslan, A., Ruman, A., Naughton, S., and Tarba, S. Y. (2021). Human dynamics of automation and digitalisation of economies: Discussion on the challenges and opportunities. In *The Palgrave handbook of corporate sustainability in the digital era*, pages 613–629. Springer.

beauty.ai (2016). The First International Beauty Contest Judged by Artificial Intelligence.

Bellamy, R. K. E. and et. al. (2018). AI Fairness

360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943 [cs]*. arXiv: 1810.01943.

Bogen, M. (2019). All the Ways Hiring Algorithms Can Introduce Bias. *Harvard Business Review*. Section: Hiring.

Carrera, F. (2020). Race and gender of aesthetics and affections: algorithmization of racism and sexism in contemporary digital image databases. *Matrizes*, 14(2):217–240.

Chiang, C. and Saw, Y. (2018). Do good looks matter when applying for jobs in the hospitality industry? *International Journal of Hospitality Management*.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. (2020). StarGAN v2: Diverse Image Synthesis for Multiple Domains. *arXiv:1912.01865 [cs]*. arXiv: 1912.01865.

Danner, M., Weber, T., Peng, L. P., Gerlach, T., Su, X., and Rätsch, M. (2021). Ethically aligned deep learning: Unbiased facial aesthetic prediction. *CoRR*, abs/2111.05149.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. ISSN: 1063-6919.

Dornaika, F., Moujahid, A., Wang, K., and Feng, X. (2020). Efficient deep discriminant embedding: Application to face beauty prediction and classification. *Engineering Applications of Artificial Intelligence*, 95:103831.

Gerlach, T., Danner, M., Peng, L., Kaminickas, A., Fei, W., and Rätsch, M. (2020). Who Loves Virtue as much as He Loves Beauty?: Deep Learning based Estimator for Aesthetics of Portraits:. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 521–528, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Grant, R. M. (1991). The resource-based theory of competitive advantage: Implications for strategy formulation. *California Management Review*.

Gupta, S. and Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science*, 161:466–474.

Gyawali, D., Pokharel, P., Chauhan, A., and Shakya, S. C. (2020). Age Range Estimation Using MTCNN and VGG-Face Model. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kärkkäinen, K. and Joo, J. (2019). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv:1908.04913 [cs]*. arXiv: 1908.04913.

Laurim, V., Arpaci, S., Prommegger, B., and Krcmar, H. (2021). Computer, whom should i hire?–acceptance criteria for artificial intelligence in the recruitment process. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 5495.

Levin, S. (2016). A beauty contest was judged by AI and the robots didn't like dark skin. Section: Technology.

Liang, L., Lin, L., Jin, L., Xie, D., and Li, M. (2018). SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction. *arXiv:1801.06345 [cs]*. arXiv: 1801.06345.

Little, A. C., Jones, B. C., and DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Lloyd, K. (2018). Bias amplification in artificial intelligence systems.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition. In *Procedings of the British Machine Vision Conference 2015*, pages 41.1–41.12, Swansea. British Machine Vision Association.

Prijatelj, D. S., McCurrie, M., and Scheirer, W. J. (2020). A Bayesian Evaluation Framework for Ground Truth-Free Visual Recognition Tasks. *arXiv:2007.06711 [cs, stat]*. arXiv: 2007.06711.

Reilly, P. (2018). The impact of artificial intelligence on the hr function.

Schlick, A. M. and Reich, K. (2021). Vereinbarkeit von akzeptanzfaktoren beim einsatz von ki in der bewerberauswahl. *Master Thesis, Reutlingen University.*

Shank, D. B. and DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, 86:401–411.

Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., and Varshney, K. R. (2020). Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 358–364, New York, NY, USA. Association for Computing Machinery.

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Xie, D., Liang, L., Jin, L., Xu, J., and Li, M. (2015). SCUT-FBP: A benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon Tong, Hong Kong, October 9-12, 2015*, pages 1821–1826. IEEE.

Zeiler, M. D. and Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*. arXiv: 1311.2901.

Zhang, M. and Kreiman, G. (2021). Beauty is in the eye of the machine. In *Nat Hum Behav 5, 675–676 ()*.