University of Tartu

Faculty of Science and Technology

Institute of Technology

Friedrich Krull

Assessment of ethnic and gender bias in automated first impression analysis

Master's Thesis (30 ECTS)

Robotics and Computer Engineering

Supervisors:

Prof. Gholamreza Anbarjafari

Kadir Aktas, MSc

Tartu 2022

Abstract/Resümee

Assessment of ethnic and gender bias in automated first impression analysis

This thesis aims to investigate possible gender and ethnic biases in state-of-the-art deep learning methods in first impression analysis. Analysing a person with some software, businesses want to find the best candidate, without the person being judged by their gender or ethnicity. To achieve this, a first impression dataset about the big five personality traits, with additional information about the person's gender and ethnic background, was used. Biases were both investigated with models trained on balanced and imbalanced data, where balanced here refers to the number of frames used from people classified as Asian, African-American, or Caucasian in the dataset. The results with both the balanced and imbalanced datasets were similar. With all the models the accuracy for Asians was much higher compared to others, which may come from the fact that the dataset did not include enough variance in the Asian data, so when evaluating, all Asians were seen similarly.

CERCS: T120 Systems Engineering, Computer Technology, T111 Imaging, Image Processing, T125 Automation, Robotics, Control Engineering

Keywords: Deep Neural Networks, Computer Vision, First Impression Analysis, Responsible AI, Human-AI Interaction

Rassilise ja etnilise eelarvamuse hindamine automatiseeritud esmamulje analüüsis

Käesoleva töö püüab uurida võimalikke soolisi ja etnilisi eelarvamusi esmamulje analüüsis kaasaegsete sügavõppe meetoditega. Analüüsides inimeste mingi tarkvaraga, soovivad firmad leida parimaid kandidaate, ilma et inimese puhul hinnatakse tema sugu või etnilist tausta. Selle saavutamiseks kasutatakse esmamulje andmestiku, mis koosneb viiest suurest iseloomujoonest, lisaks on infot isikute soolisest ja etnilisest taustast. Eelarvamusi uuriti nii mudelitega, mis olid treenitud tasakaalustatud ja tasakaalustamata andmete peal, siinkohal näitab tasakaalustatus, kasutatud kaadrite arvu treenimiseks olenevalt inimeste etnilisest taustast andmestikus. Tulemused olid sarnased nii tasakaalustatud, kui ka tasakaalustamata andmestikega. Kõigi mudelite puhul ilmnes, et asiaatide täpsus oli teistega võrreldes kõrgem, mis võis tuleneda sellest, et asiaatide andmetes puudus piisav dispersioon ning selle tõttu nähti kõiki asiaate sarnastena.

CERCS: T120 Süsteemitehnoloogia, Arvutitehnoloogia, T111 Pilditehnika, T125 Automatiseerimine, Robootika, Juhtimistehnika

Märksõnad: Sügavad Närvivõrgud, Tehisnägemine, Esmamulje Analüüs, Vastutustundlik Tehisintellekt, Inimese-Tehisintellekti Interaktsioon

Contents

Abstract/Resümee **2**

List of Figures **6**

List of Tables **8**

Abbreviations **9**

Introduction **10**

2 Literature review **12**

2.1 Big Five Personality Traits 12

2.2 Bias in Deep Neural Networks 13

2.3 Deep learning in computer vision 13

2.4 Convolutional Neural Networks 14

2.4.1 Convolutional Layer 14

2.4.2 Pooling Layer 16

2.4.3 Fully-connected Layer 16

3 Methodology **18**

3.1 Baseline model 18

3.2 Modified DAN model 18

3.3 Modified DAN model only face 19

3.4 Modified VGG-16 20

3.5 VGG-16 20

4 Dataset **21**

4.1 ChaLearn FI dataset 21

4.2 Data Exploration 21

4.3 Data Preprocessing 23

4.4 Balanced Data 25

4.5 Model Evaluation 25

4.5.1 Hardware 26

4.5.2 Software 26

5 Experimental Results **27**

5.1 Baseline 27

5.2 Modified DAN 28

5.3 Modified DAN trained on the face 28

5.4 Modified VGG-16 29

5.5 VGG-16 30

5.6 Balanced models 30

6 Analysis **32**

6.1 Gender 32

6.2 Ethnicity 35

6.3 Balanced 39

6.3.1 Gender 39

6.3.2 Ethnicity 39

6.4 Comparison 40

Summary **42**

Bibliography **45**

Non-exclusive licence **49**

List of Figures

1 Convolution with a 3×3 kernel and stride one. 15

2 An example of a 3×3 kernel where the stride is three. 15

3 An example of using either max or average pooling with a 2×2 filter and with a stride of 2 when conducting pooling. 16

4 The original architecture of ResNet-18 [26]. 19

5 The network architecture that is based on the DAN+ architecture that won second place in the 2016 edition of the ChaLearn competition [28]. 19

6 The structure of the VGG-16 model. 20

7 The distribution of values for each personality trait and interview in every video in the training dataset. 23

8 Frame from a video, with its personality and interview scores. 23

9 Video frame with only the facial region cut out. 24

10 The mean predicted value for openness for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 32

11 The mean predicted value for conscientiousness for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 33

12 The mean predicted value for extraversion for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 34

13 The mean predicted value for agreeableness for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 34

14 The mean predicted value for neuroticism for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 35

15 The mean predicted value for openness for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 36

16 The mean predicted value for conscientiousness for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 36

17 The mean predicted value for extraversion for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 37

18 The mean predicted value for agreeableness for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 38

19 The mean predicted value for neuroticism for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth. 38

20 The mean predicted values for openness, extraversion, agreeableness and neuroticism compared to the ground truth for men and women in the balanced dataset. 39

21 The mean predicted values for openness, extraversion, agreeableness and neuroticism compared to the ground truth for Asians, African-Americans, and Caucasians in the balanced dataset. 40

List of Tables

1 Distribution of ethnicity and gender in the datasets. 22

2 Gender distribution among ethnicities. 22

3 Frame counts for each group used for training models on the balanced data. . . 25

4 Mean accuracies of different groups in the baseline model. 27

5 Mean accuracies of different groups with the modified DAN model, where the input is the original frame. 28

6 Mean accuracies of different groups in the modified DAN model, where the input is the processed face. 29

7 Mean accuracies of different groups with the modified VGG-16 model. 29

8 Mean accuracies of different groups with the ResNet-18 model trained on the balanced dataset. 30

9 Mean accuracies of different groups with the modified DAN model trained on the balanced dataset. 31

10 Mean accuracies of different groups with the modified DAN model trained on facial region and balanced dataset. 31

11 Mean accuracies of different groups with different methods used. 41

Abbreviations

AI - Artificial Intelligence

CAM - Class Activation Maps

CNN - Convolutional Neural Network

CPU - Central Processing Unit

DAN - Descriptor Aggregation Networks

DNN - Deep Neural Network

FC - Fully Connected

FFM - Five-Factor Model

FI - First Impressions

GPU - Graphics Processing Unit

OCEAN - Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

RGB - Red, Green, Blue

Introduction

Large companies receive a lot of job applications, enough for it to be impossible to process all of them properly. For this reason, companies try to look for alternative methods, where some software does the filtering for them, and personnel only looks at the top of the top applicants. Today, almost all Fortune 500 companies already use some systems to aid human resources in the hiring process [1].

As technology improves and recording devices become more available to people, it is possible to conduct video interviews where the candidates record themselves answering predefined questions without the need to have a face-to-face meeting with everybody. Now a new issue arises on how these videos will be evaluated because sitting through all these interviews would possibly be even more time-consuming compared to just checking the resumes.

This is where computer vision can come in, computers are capable of processing huge amounts of data and can run day and night but here a question arises on how to tell a computer to determine, which interviewees have the right personality traits for the job at hand. To solve this problem, huge steps have been made in the field of deep learning, where it is possible to train deep neural networks to sort out the candidates, with lower personality scores. To complete this, neural networks would however need data, which could then be used to train models and determine the personality of the interviewee [2]. Neural networks have shown to give great results, but one of the issues surrounding them is that they are tough to explain and can be seen as ”black boxes” [3]. These ”black boxes” are often used as part of an AI, and the need for data here does not always take into account the distribution of different people in the real world thus, it could lead to biases towards some groups [4]. With the use of these AIs, there is often an issue that problems regarding unfair treatment of people are discovered only after deployment, which causes inconveniences for people using these [5, 6]. For that reason, it is necessary to investigate possible avenues of unfairness in deep neural networks.

Problem overview

As already mentioned, companies are looking for ways to automate the search for newer employees, and that neural networks could be the way to go however with neural networks being ”black boxes”, a question comes up, how is it possible to determine whether the network evaluates the personality traits equally for everyone, irrespective of their gender or ethnic background. Previously, Principi *et al.* have investigated, how utilizing different human attributes such as age, gender, ethnicity etc could affect the predictions. They found that with the inclusion of gender, the results favoured women [7]. In addition, Escalante *et al.* have investigated gender and ethnicity biases in the first impressions dataset and found that the data itself seemed to be positively biased towards females and that the ethnic bias was weaker compared to the gender bias but the results indicated that the data was positively biased towards Caucasians, while African-Americans were seen more negatively. They did not detect any discernible bias towards Asians [8].

Goals

The features that are not controllable by the person such as gender, and ethnicity should not play a key role in how people are perceived. This paper aims to assess the state-of-the-art DNN architectures are used in first impression analysis and to investigate possible avenues of bias in gender or ethnicity.

To achieve this goal, the first task is to train models using the first impressions dataset and to investigate, how accurate the method is for the different groups. In addition, since the dataset itself does not contain equal amounts of each group, the next idea is to use a more balanced dataset, where the number of frames for each group is relatively more similar compared to the default distribution. Finally, using these points to determine, if there is an underlying theme of bias among the results.

2 Literature review

The state-of-the-art study into first impression analysis uses deep learning to analyse the big five personality traits. Therefore, firstly there will be a brief overview of what the big five are. After that, as the big five are based on human characteristics and here these are used in deep learning, then a brief overview of bias in DNN is given. Finally, an overview of deep learning is given that has relevance to personality analysis.

2.1 Big Five Personality Traits

Humans get a lot of information just from seeing another person, the way a person tilts or turns their head can affect how someone else is perceived. Behavioural psychologists have tried to understand this in humans for a while now, and multiple psychological models have been brought up. The model that has found the most widespread use has been the Big Five model [9].

The Big Five or Five-Factor Model (FFM) itself, as the name would suggest, uses five different human characteristics to determine the overall personality of each being. The factors that have most often been used to characterize a human's personality this way are openness, which shows how interested a person is to in ideas or experiences, and conscientiousness, which shows how willing a person is to strive to achieve their goals and how far they wish to go in the future, extraversion, which shows the person's willingness to communicate with other people and their overall social skills, agreeableness, which is used to show how forgiving a person is and how they seem to cope with other people's ideas that are not in line with their own, neuroticism, which shows the person's feelings of self-doubt. These five factors are collectively often referred to as simply OCEAN, which arises from the first letter of each factor. These factors are openness, conscientiousness, extraversion, agreeableness, and neuroticism [10].

2.2 Bias in Deep Neural Networks

People are different, but when interacting with others for the first time, we often make a split-second decision of whether this person is someone we may continue communicating with or someone not suitable to us, this can happen subconsciously without us even noticing [11]. When deciding on whether the person we are currently interacting with is of interest then it needs to be also said that the initial judgement also depends on the person who is judging since how much a mannerism affects a person depends on how important, they themselves find it, which could lead to one person being judged much higher by one person compared to another [12].

Given that humans have biases towards different groups, as brought out before, this brings up a new question, how could these biases apply to deep neural networks (DNN). Most of the labelled data used in personality analysis has been initially labelled by people, and there has already been an occasion, where AI has shown signs of being biased towards one ethnicity or gender [13]. Research into the field of biases in DNN is ongoing and involves a wider range of classification and regression tasks related to people [14]. It has been found that a bias factor in gender classification is age, as older people are more likely to be classified correctly [15]. The main study that has gone into investigating bias in the FI dataset for the big five is by Principi *et al.*, who investigated bias in a multimodal system, utilizing both the auditory and visual data as well as using pre-processed faces to predict age, gender, ethnicity etc and using information collected from these to get the final big five values [7].

2.3 Deep learning in computer vision

Deep learning makes it possible to create computational models which can include multiple layers to represent and process data similarly to how the human brain computes it. Deep learning itself includes a wide array of methods such as neural networks, hierarchical probabilistic models, as well as supervised and unsupervised learning algorithms. Deep learning has been the main focus of interest because of its ability to exceed the performance of previous state-of-the-art methods in a multitude of tasks, and its ability to understand complex data, such as visual and auditory data [2, 16–18].

Some of the factors that have attributed to the continued rise of deep learning are the rapidly growing amounts of publicly available high-quality labelled datasets, which give people easier

access to previously unavailable data as well as the rise of parallel GPU computing, which has allowed to transition from CPU-based training to GPU-based, which in turn has significantly sped up the training process. The availability of publicly available frameworks like TensorFlow, which make it much easier for people to create networks for model training, has also greatly contributed to the rise of deep learning [2].

2.4 Convolutional Neural Networks

Convolutional Neural Network (CNN) is a framework that has found wide use in deep learning and has often been used for object recognition tasks. The initial inspiration behind CNNs was the neocognitron, which was published in 1980 [19] and later improved in 1990 with the LeNet [20]. This method was mainly used for determining handwritten digits without prior preprocessing, but it failed with more difficult problems, as there was not enough training data or computing power available. The next big step was made in 2012 with the AlexNet [21] which won the 2012 edition of the ImageNet Large Scale Visual Recognition Challenge [22].

CNNs are similar to regular neural networks but the main difference between the two is that in the CNN the hidden layer neuron has connections only to a subset of neurons in the previous layer. The sparseness of the connections allows for this network to learn features implicitly and since the whole architecture of the network is deep, it allows extracting different parts of the features in different layers for example extracting the edges of the objects first and then after knowing the edges it is possible to determine the overall shape of the object and in the end with all the knowledge of how the object itself looks like, it is possible to determine what the actual object represented is [22].

2.4.1 Convolutional Layer

The convolutional layer is the most important part of CNNs. This layer is where most of the computations are carried out [22]. CNNs are mostly used when dealing with images and when using all the pixels for one image as input then without using convolution the number of weight connections required would become colossal, rapidly. This is why it was thought that only using regional connections to give information to the next layer would be better. In addition to using only regional connections, the local connection weights are also fixed for all the neurons in the next layer. This means that the next layer's neighbour neurons are connected without

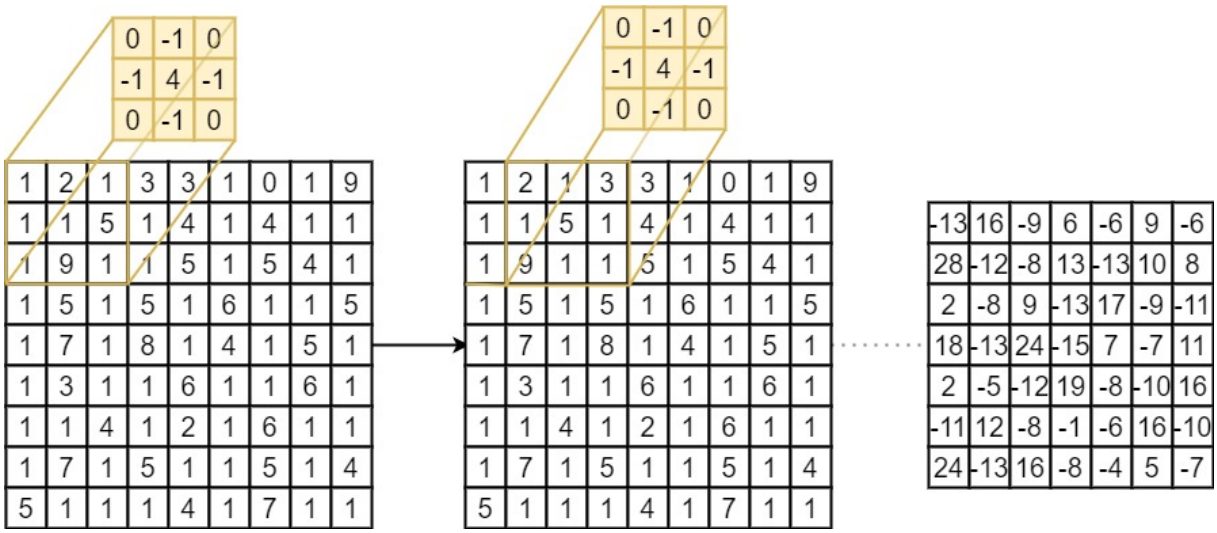


Figure 1. Convolution with a 3×3 kernel and stride one.

changing the weight of the local region in the previous layer. This also reduces the number of connections. These small changes to the network allow for the number of connections to massively decrease. In addition, fixing the weights for local connections results in a similar effect of having a window slide over the input and after that mapping the freshly produced output in place of the input. The effect of sliding over the image is why this method is called convolutional, as can be seen in Figure 1, and the convolutive effect it produces allows the method to find and determine features on an image despite the object's location. To increase the performance of this method, it is possible to combine multiple layers and thus each layer could function as a different filter, which results in the possibility of extracting multiple features from the same image [23].

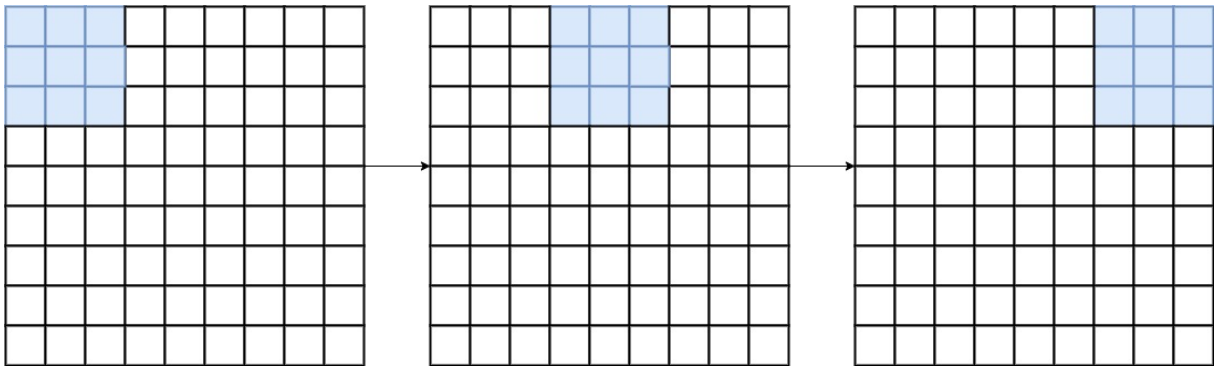


Figure 2. An example of a 3×3 kernel where the stride is three.

To lessen the number of parameters even further, one option is to use the stride parameter, which allows specifying the step to be taken when doing convolutions on a layer, an example of stride is shown in Figure 2. If the wish is to process the centre of the image then usually

choosing smaller values for the stride is more beneficiary but selecting larger values is better suited towards processing the edges [24].

Stride allows specifying how large steps are taken when looking at the current layer, but this could lead to a loss of information on the edges as with large filters these values are never reached however by adding zero-padding the input image is padded with values and thus allows to gather information about pixels located on the border of the image. Padding also allows manipulating the dimensions of the output layer [23].

2.4.2 Pooling Layer

Pooling is essential in networks that use convolution, as it decreases the dimensionality of the feature map. It takes a larger image, for example, and lessens its size. The main purposes of this layer are to lower the number of parameters required, as well as to control overfitting. The two simplest pooling methods used are average and max pooling. Average pooling, as the name states, looks at all the values in the given filter's area and takes the mean value over all pixels and max pooling takes the maximal value in the chosen area, an example of this is shown in Figure 3 [25].

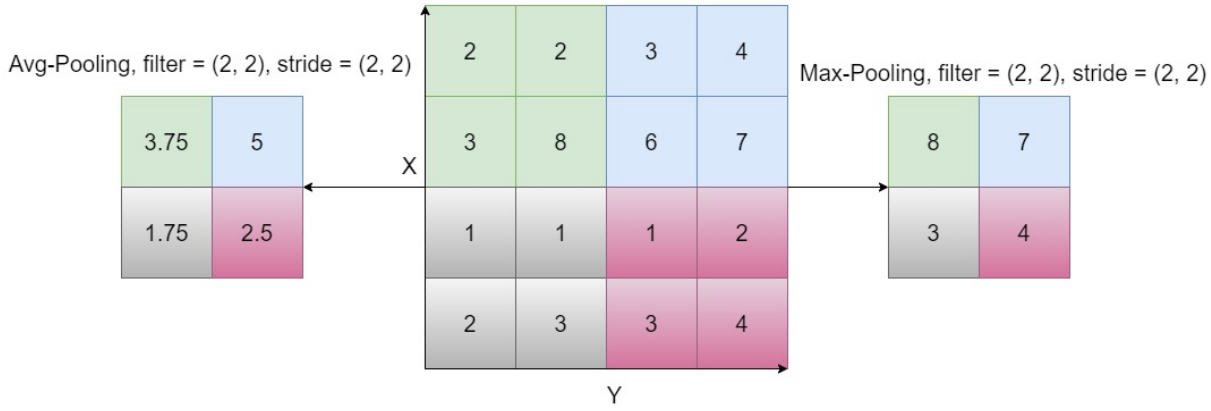


Figure 3. An example of using either max or average pooling with a 2×2 filter and with a stride of 2 when conducting pooling.

2.4.3 Fully-connected Layer

The fully-connected (FC) layer is essentially the same in functionality as the neuron arrangement in traditional neural networks. This means that all the nodes in the previous layer are

connected to all the nodes in the next layer. The downside of the FC layer is that it requires lots of parameters, and it performs complex computations in training [23].

3 Methodology

Technological advancements in deep learning-based methods have found success in many computer vision tasks. Consequently, its techniques have found uses in first impression analysis as well. State-of-the-art results have been achieved with CNN based deep architectures like ResNet and VGG. As this thesis focused on the visual information then, similar to Escalante *et al.* where the visual baseline for accuracy was the ResNet-18 structure, it was used here as well [8]. VGG structure has also been used in a variety of tasks relating to images, usually mainly aimed at image classifications. Moreover, the DAN+ structure has achieved state-of-the-art results, and it has previously been modified for use in first impression analysis for visual data.

3.1 Baseline model

To get a first look into how the biases in the models could affect different genders and ethnicities, an initial model was chosen that could be used as a baseline and can be compared to other methods. The selected model has a ResNet-18 structure (shown in Figure 4). The last layer of this model was changed so that it would output six labels, which each correspond to the different Big Five personality traits, as well as the interview. The training process included using frames from all the 6,000 training videos and each frame from the same video was given the same value, which corresponded to the value in the labels table [8].

3.2 Modified DAN model

After the baseline, a secondary model was chosen. The second model was chosen, so that the reported accuracy would be higher compared to the baseline, as well as only utilizing the visual data to predict the features. The selected model is based on DAN+, which itself is an

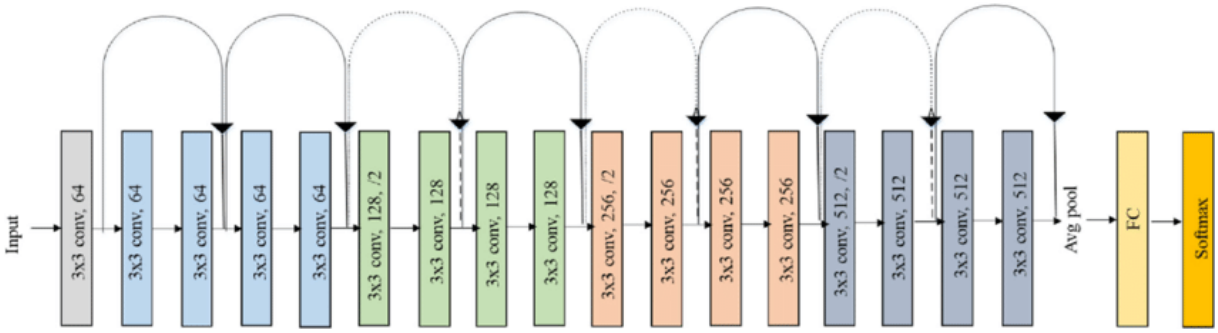


Figure 4. The original architecture of ResNet-18 [26].

extension to the DAN model, this model itself was used by the team that placed second in the 2016 edition of the ChaLearn competition [27], where layers after average and max pooling after pool5 (shown in Figure 5) with the Class Activation Map module, in addition, the average pooling and the max pooling after relu5_2, which were used in the DAN+ model were removed. The model was trained with ten frames from each video, which were given the same trait values as the full video [28].

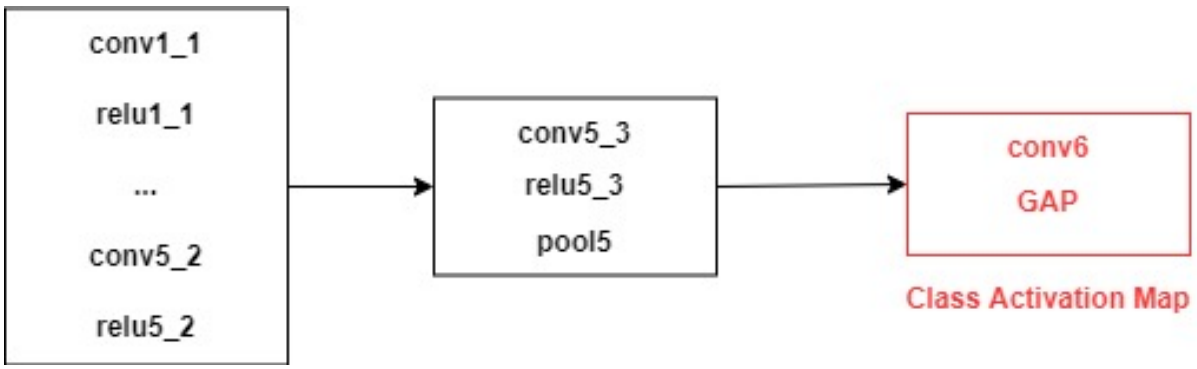


Figure 5. The network architecture that is based on the DAN+ architecture that won second place in the 2016 edition of the ChaLearn competition [28].

3.3 Modified DAN model only face

Ventura *et al.* found after looking into the CAMs that the highest support areas overlap with where the person's face is situated in the frame, as such they secondly used only the facial part of the frame. The architecture of the model and the number of frames used per video stayed the same, but instead of a $224 \times 224\,px$ frame, a $224 \times 224\,px$ facial region was used [28].

3.4 Modified VGG-16

The VGG-16 (structure shown in Figure 6) was created for image classification tasks [29], but by modifying the structure of the model, it is possible to make it usable in regression tasks. Here, the VGG-16 model has been modified similarly to Helm *et al.* [30] where they modified the activation function of the last FC layer to be the sigmoid function and added a Batch-Normalizations after convolutional layers and FC layers (except for the output layer). Instead of the mean squared error loss function, the one used in this thesis was mean absolute error. For training, only the extracted face was used.

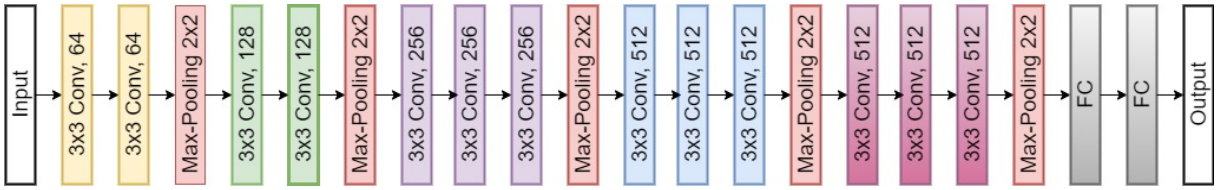


Figure 6. The structure of the VGG-16 model.

3.5 VGG-16

In addition, to using the modified VGG-16 model where Batch-Normalizations were added after each convolutional layer, the regular structure of the VGG-16 model was also used for training, the output layer was still modified here so that it could output the Big-Five features.

4 Dataset

The dataset that was used for training, validating and testing the model is called ChaLearn First Impressions (FI) dataset [31]. In addition, an extra table was included where for each video there was information about the gender of the person being male or female as well as the ethnicity of the person. Ethnicities were separated into three major groups, Asian, African-American and Caucasian descent.

4.1 ChaLearn FI dataset

The ChaLearn FI dataset consists of 10,000 video clips that have been extracted from high-definition YouTube videos, the length of these videos is mostly around 15 seconds. The selected videos were filtered so that each video would only include one person, who is over the ages of 13-15 and was speaking English. The dataset contains people of different genders, ages, nationalities and ethnicities. The 10,000 clips are separated into training, validation and testing clips with a ratio of 3:1:1, which means that 6,000 clips are used for training, 2,000 for validating and another 2,000 for testing. The video labels include the big five personality traits and an additional label interview which shows how likely this person is to be invited to a job interview. The trait values are continuous in the range $[0, 1]$. Since these videos are labelled by people, then there is a possibility of bias from a person towards another person. These biases may come from evaluators' prejudices towards some ethnicity, age group or gender [31].

4.2 Data Exploration

When determining the biases in the final models, it would be preferable to see how the differences in the initial training data could shape the resulting models' output, so it is good to see what the data that is being used of consists. The gender and ethnicity amounts are brought in

Table 1. Distribution of ethnicity and gender in the datasets.

Dataset	**Videos**	**Males**	**Females**	**Asian**	**African-American**	**Caucasian**
Training	6000	2734	3266	215	623	5162
Validation	2000	916	1084	68	224	1708
Test	2000	888	1112	48	224	1728

Tables 1 and 2. It is possible to see from Table 1 that the number of females and males in videos is quite similar but the gender distribution is very imbalanced as most of the dataset consists of Caucasians.

Table 2. Gender distribution among ethnicities.

Gender and Ethnicity	**Training set**	**Validation set**	**Test set**
Asian Male	64	18	14
Asian Female	151	50	34
African-American Male	201	67	70
African-American Female	422	157	154
Caucasian Male	2469	831	804
Caucasian Female	2693	877	924

When taking into account the ethnicity and gender of the people that are represented in the videos then Table 2 shows that people of Caucasian descent make up about $\frac{5}{6}$ of the data and the number of both genders is about equal but for the other ethnicities the numbers are very small in comparison and as such one problem that could arise is that this dataset if taken by default may be imbalanced. This imbalance could lead to the models being biased toward something that is in the majority in the whole dataset, thus resulting in poorer performance for those that are in the minority [32].

Another thing which could lead to imbalanced datasets is the distribution of the labelled data, which is shown in Figure 7. The distribution itself resembles a Gaussian distribution, where values that belong to the centre have more examples in the dataset and the edge cases are very rare.

In addition, as stated before for each video in the dataset there is a corresponding value for the

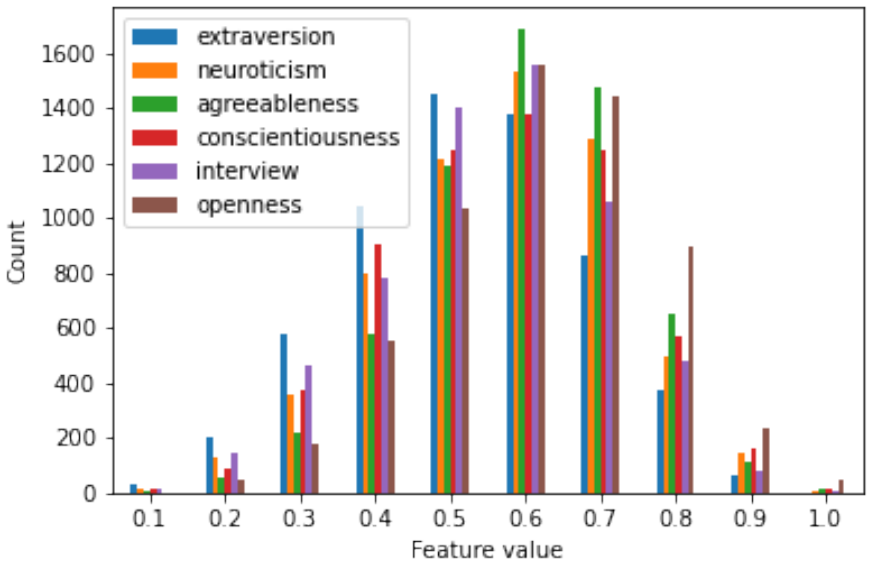
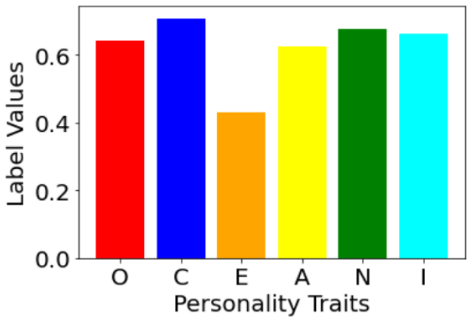


Figure 7. The distribution of values for each personality trait and interview in every video in the training dataset.

five major personality traits as well as the value, which indicates how likely this would be called up to an interview, this can be seen in Figures 8a, and 8b.

(a) Frame from a video in the training set.



(b) Labels that correspond to the video that the frame is taken from.

Figure 8. Frame from a video, with its personality and interview scores.

4.3 Data Preprocessing

Firstly, as all the data was zipped together the initial task was to unzip the files for this a slightly modified version of the unzipping method used by Zhang *et al.* was deployed, which separated the data into videos [27].

Figure 9. Video frame with only the facial region cut out.

After the video files were extracted, the videos were separated into frames using OpenCV and during preprocessing each frame was resized to $456 \times 256\ px$ as these frames would later be resized again, then the main part of this resizing was to save disk space, here the acronym px represents pixels [8]. Additionally, from all the video clips in the dataset, the person's face with dimensions of $224 \times 224\ px$ was extracted using OpenFace 2.0, which is a toolkit that is the extension to the original OpenFace that has higher accuracy in detecting facial landmarks and can estimate the pose of the head and the eye-gaze as well as recognizing facial action units [33]. The faces that get cut out from the initial video frames are shown in Figure 9. In addition, since the input data is often normalized, so that the values got from pixels in the image would be more similar [34]. The initial data has values in the range $[0, 255]$ but after applying Min-Max Normalization all the data is in a range of $[0, 1]$. The equation for the Min-Max Normalization, x_{norm} is as follows:

$$x_{norm} = \frac{x - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where x is the pixel being normalized, X_{min} is the lowest pixel value in the whole dataset and X_{max} is the highest value in the whole dataset [35].

4.4 Balanced Data

The models that were used for training were later retrained so that the count of frames for different gender and ethnicity groups would be similar and to see how this would affect the potential biases towards different groups. This was implemented as when looking at Table 1, it is possible to see that the majority of the data is made up of people of Caucasian descent, Asians and African-Americans make up a smaller minority in this dataset. The number of females and males overall in this dataset however is quite similar. After extracting a similar number of frames for each group, the number of frames for each group, used for training, is shown in Table 3.

Table 3. Frame counts for each group used for training models on the balanced data.

Dataset	**Videos**	**Males**	**Females**	**Asian**	**African-American**	**Caucasian**
Training	168819	83801	85018	53978	55590	59251
Validation	55805	26676	29129	16578	19608	19619

4.5 Model Evaluation

All the labels have a continuous value in the range $[0, 1]$, which means that this is seen as a regression problem as such the metric used for accuracy is mean absolute accuracy, which for each of the big five traits can be found as:

$$A = 1 - \frac{1}{N}\sum_{i=1}^{N}|t_i - p_i| \tag{2}$$

where A represents the accuracy for each trait, N is the number of total videos, t_i is the actual value of the trait, p_i is the predicted value for the video [31]. The higher the value, the better the accuracy of the video, and the accuracy of each video was taken as the mean over all the frames in the video.

In addition, when comparing the accuracies of different groups a relative difference is used to find the difference between the two values:

$$d = \frac{|x - y|}{\frac{|x+y|}{2}} \tag{3}$$

where the d stands for the difference between two items, x is the accuracy of one group and y is the accuracy of another group [36].

4.5.1 Hardware

The specifications of the computer that was used to train the models used in the thesis:

- Processor: Intel® Core™ i7-11800H 2.3 GHz.
- Installed Memory(RAM): 8.00 GB.
- Graphics Card (GPU): NVIDIA GeForce RTX 3050 4 GB.

4.5.2 Software

The coding and training were done using Python 3.8 and in addition, the following packages and tools were used in data processing and training:

- Pandas 1.2.4
- Pillow 8.2.0
- OpenCV 4.5.4.60
- TensorFlow 2.7.0
- NumPy 1.20.1
- Matplotlib 3.3.4
- OpenFace 2.2.0
- NVIDIA cuDNN 11.5

5 Experimental Results

To get more information about how the models are fair, their accuracy was found using Equation 2. After extracting the accuracy for each test video, the data was also separated into both the gender and ethnic groups to get the mean accuracy for all. This, in addition, allowed us to compare the different groups and see, whether there is bias or not.

5.1 Baseline

Table 4. Mean accuracies of different groups in the baseline model.

Group	O	C	E	A	N	Average
Overall	0.89926	0.90381	0.89986	0.90335	0.8974	0.90074
Male	0.90137	0.90121	0.90225	0.90112	0.90074	0.90134
Female	0.89758	0.9059	0.89795	0.90513	0.89473	0.90026
Asian	0.92188	0.92421	0.92183	0.92013	0.92167	0.92194
Caucasian	0.89943	0.90374	0.89978	0.90364	0.89659	0.90063
African-American	0.89314	0.89999	0.89575	0.89756	0.89843	0.89697

The training of this model included the interview label, however, as this was omitted in other models then it was not included in the results section. The initial model had an overall average absolute accuracy of, 0.90074 without the interview label, for which all the different features with their accuracies are brought out in Table 4. From Table 4 it is possible to see that most accuracies are between 0.89 and, 0.90; however, for people who belong to the Asian ethnicity group their accuracy is a bit higher. Comparing the accuracies of different groups to the overall accuracy, the difference in accuracy between men and women is small where they differ from the overall by 0.0007 and 0.0005 respectively. When comparing the overall average accuracy to

different ethnicities, then the relative difference in accuracy is for Caucasians 0.0001, African-Americans 0.0042 and Asians 0.0233. The relative difference is found using Equation 3

5.2 Modified DAN

Table 5. Mean accuracies of different groups with the modified DAN model, where the input is the original frame.

Group	O	C	E	A	N	Average
Overall	0.9048	0.91233	0.90585	0.90538	0.90187	0.90605
Male	0.90394	0.90969	0.90573	0.90368	0.90275	0.90516
Female	0.90549	0.91443	0.90594	0.90673	0.90117	0.90675
Asian	0.93084	0.9248	0.921	0.92124	0.93431	0.92644
Caucasian	0.90509	0.91224	0.90602	0.90539	0.90141	0.90603
African-American	0.89699	0.91035	0.90128	0.90187	0.89844	0.90179

The model that was based on the modified DAN model had an overall accuracy of 0.90605 and did not include the interview label in the model structure, as leaving this out when finding the Big-Five features will lessen the risk of implementing bias into this feature. The accuracies of different groups can be seen in Table 5, where similarly to the ResNet-18 model, the highest mean accuracy belongs to the Asian group. The relative difference between males and females when compared to the overall is 0.001 and 0.0008 respectively and when looking at the different ethnic groups the biggest difference from the overall average accuracy is with Asians, where it is 0.0223 and the African-American group has a difference of 0.0047 the Caucasian group was very close to the actual prediction accuracy and the difference was only 0.00002.

5.3 Modified DAN trained on the face

The model that was based on the modified DAN model and only used the detected facial region had an overall accuracy of 0.91109 and did not include the interview label in the model structure, as leaving this out when finding the Big-Five features will lessen the risk of implementing bias into this feature. The accuracies of different groups can be seen in Table 6.

Table 6. Mean accuracies of different groups in the modified DAN model, where the input is the processed face.

Group	O	C	E	A	N	Average
Overall	0.9084	0.9161	0.91374	0.90936	0.90782	0.91109
Male	0.90651	0.91229	0.91286	0.90758	0.90787	0.90942
Female	0.9099	0.91914	0.91445	0.91079	0.90779	0.91241
Asian	0.92647	0.93317	0.9298	0.91595	0.92783	0.92664
Caucasian	0.90767	0.91582	0.91352	0.90898	0.90708	0.91061
African-American	0.91013	0.91462	0.91199	0.91089	0.90931	0.91139

Comparing this to previous models, the relative accuracy difference for the Asian group is highest in this case as well, with 0.0169 but compared to before this is the lowest difference so far. The absolute difference between males and females is the highest so far with 0.0018 and 0.0015. The Caucasians and African-Americans overall have the closest absolute accuracy difference to the overall, the differences being 0.0005 and, 0.0003 respectively.

5.4 Modified VGG-16

Table 7. Mean accuracies of different groups with the modified VGG-16 model.

Group	O	C	E	A	N	Average
Overall	0.87341	0.85515	0.86906	0.87427	0.84461	0.8633
Male	0.87137	0.85737	0.87302	0.88076	0.84078	0.86466
Female	0.87505	0.85338	0.86589	0.86908	0.84767	0.86222
Asian	0.91385	0.89046	0.89552	0.88611	0.8854	0.89427
Caucasian	0.87339	0.85267	0.86802	0.87328	0.84343	0.86216
African-American	0.86496	0.86677	0.87136	0.87934	0.84499	0.86548

The model that had the VGG-16 model modified to include Batch-Normalization after every convolutional layer had an overall accuracy of 0.8633 and, similarly to the modified DAN model, this model's output was the Big-Five traits. The accuracies achieved with this model

are brought out in Table 7. The relative difference between males and females was 0.0016 and 0.0013 respectively, and for the ethnicity groups the relative differences were for Asians 0.0352, which once more is the largest and for Caucasians and African-Americans the differences were 0.0013 and 0.0025.

5.5 VGG-16

The model that used the structure of the VGG-16 failed, as it was not able to predict anything and the model's output was always the same, thus resulting in it not working correctly and making it not possible to give any accurate results.

5.6 Balanced models

The tests that were carried out on the dataset that included similar amounts of inputs for both genders and each of the three ethnicities ended up showing similar results, which have been brought out in Tables 8, 9, and 10.

Table 8. Mean accuracies of different groups with the ResNet-18 model trained on the balanced dataset.

Group	O	C	E	A	N	Average
Overall	0.87358	0.86899	0.85855	0.88407	0.86423	0.86988
Male	0.87732	0.86481	0.85155	0.88375	0.87035	0.86956
Female	0.87059	0.87232	0.86414	0.88433	0.85934	0.87015
Asian	0.91203	0.91028	0.90477	0.90532	0.89778	0.90604
Caucasian	0.87338	0.86676	0.85767	0.88254	0.86224	0.86852
African-American	0.86693	0.87728	0.85539	0.8913	0.8724	0.87266

These show the results from the baseline and both of the modified DAN variants. The results from these look almost identical to the ones received from the models trained on the imbalanced dataset, except for the modified DAN, which was trained on the facial region and the reason for that is, that it was unable to predict conscientiousness. Every prediction with that model gave a conscientiousness of zero, which meant its accuracy was much lower compared to the other

Table 9. Mean accuracies of different groups with the modified DAN model trained on the balanced dataset.

Group	O	C	E	A	N	Average
Overall	0.90458	0.90977	0.90453	0.90489	0.90012	0.90478
Male	0.90574	0.90782	0.90611	0.90412	0.90304	0.90536
Female	0.90365	0.91133	0.90328	0.90551	0.89779	0.90431
Asian	0.93321	0.93095	0.92307	0.92374	0.93518	0.92923
Caucasian	0.90486	0.90959	0.90462	0.90497	0.89972	0.90475
African-American	0.8963	0.90658	0.8999	0.90025	0.89569	0.89974

methods. Out of the three balanced models trained, modified DAN worked the most similar to the imbalanced version but the accuracy of both the baseline and the facial modified DAN model worsened if the conscientiousness were to be discarded then this was also very similar to the original version.

Table 10. Mean accuracies of different groups with the modified DAN model trained on facial region and balanced dataset.

Group	O	C	E	A	N	Average
Overall	0.90755	0.47485	0.91033	0.90752	0.90378	0.82081
Male	0.90512	0.48977	0.91033	0.90447	0.90205	0.82235
Female	0.90949	0.46295	0.91032	0.90996	0.90515	0.81957
Asian	0.9314	0.46299	0.92607	0.91684	0.9367	0.8348
Caucasian	0.90792	0.47008	0.91092	0.90758	0.90362	0.82002
African-American	0.89963	0.51426	0.90238	0.90506	0.89793	0.82385

6 Analysis

6.1 Gender

When taking into account the information that is brought out in Tables 4, 5, 6, and 7, it is possible to compare the accuracies for males and females. The results that are gained from the difference in female and male predictions should best represent the whole dataset, as the number of males and females in the dataset is quite similar.

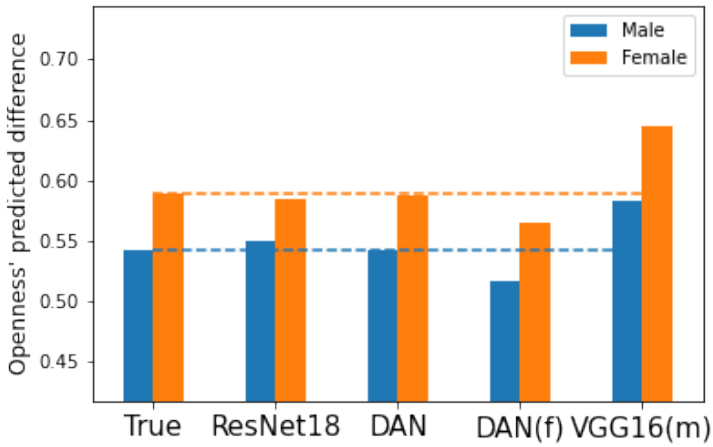


Figure 10. The mean predicted value for openness for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

The first trait of the Big-Five that will be compared here is the openness trait. Looking at the results from the four models that gave results which can be interpreted, then there is no trend for all the models that would say that one gender is always predicted more accurately, the ResNet-18 predicts the results better for men compared to women and the modified VGG-16 model with Batch-normalizations and the models based on the modified DAN structure are more accurate towards women. Looking at the average predicted value for openness compared to the actual

value, which is visible in Figure 10, looking at this figure it is possible to see that the initial baseline model, which is shown as ResNet-18, rated males higher compared to the actual value and females lower, with the modified DAN model that was trained on the full-frame both of the predictions were similar to the actual values on average however it slightly favoured men and predicted their results higher than the actual value. In the modified VGG-16 model and the modified DAN model that was trained on only the facial region, the results are the opposite of each other. The modified DAN with facial input evaluates both males and females as worse than the actual value, and the modified VGG-16 evaluates them as better.

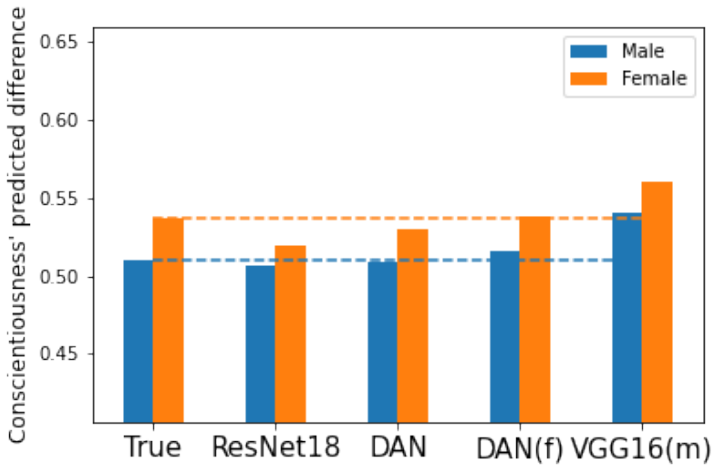


Figure 11. The mean predicted value for conscientiousness for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

Moving onto the other traits in the OCEAN, I will next look at how conscientiousness compares with different models. Again, when looking at the mean accuracies for both males and females similar to the openness trait, one is not definitively better in all variants. The accuracy for females is higher in the ResNet-18, and both DAN versions, but with the modified VGG-16 the accuracy for females is a bit worse. Looking at how these models' predictions compare to the actual values (shown in Figure 11) then the ResNet-18 model predicted males as having much better conscientiousness than they actually did and for women, it was predicted as a bit worse. With conscientiousness, one trend does arise that with all models compared to the actual value, males are seen as more conscientious.

The accuracy of extraversion in different models is similar to the previously looked at traits, as there is no clear indication with it to say that something is always predicted more accurately compared to the other.

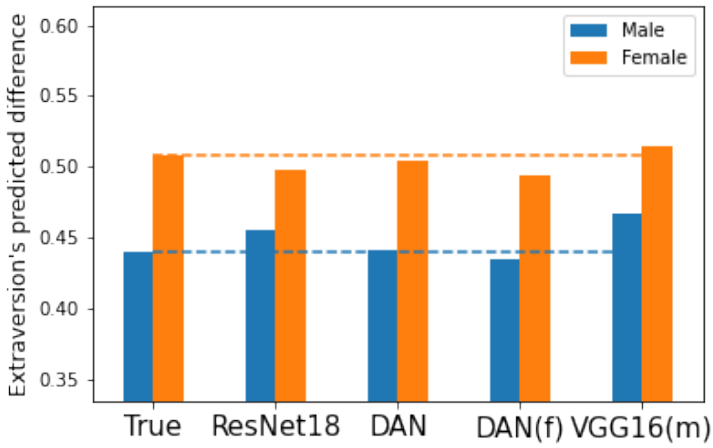


Figure 12. The mean predicted value for extraversion for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

The ResNet-18 and the modified VGG-16 models gave higher accuracies to men, but both variants of the modified DAN model favour the women, but when looking at the overall trend of extraversion in different models for both genders as can be seen in Figure 12 then it is possible to see that for all the cases when comparing it to the actual results men are predicted to be more outgoing than they actually are and women when comparing to how the men predicted to behave are predicted to be less outgoing than in reality.

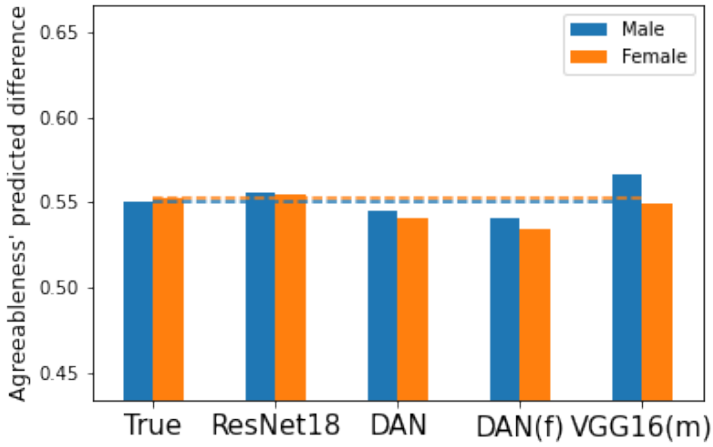


Figure 13. The mean predicted value for agreeableness for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

The accuracies with different models for the agreeableness give similar results to other traits, in that with ResNet-18, and the modified DAN variants females have a better accuracy but with

the modified VGG-16 model the accuracy for men is much higher. Now when comparing the average predicted values for agreeableness to the average actual agreeableness then since this is the closest the actual values have been to each other, then this is the first time when predicted values for men have been higher compared to women. In every model, it is possible to see that predictions ended up favouring men, as can be seen in Figure 13.

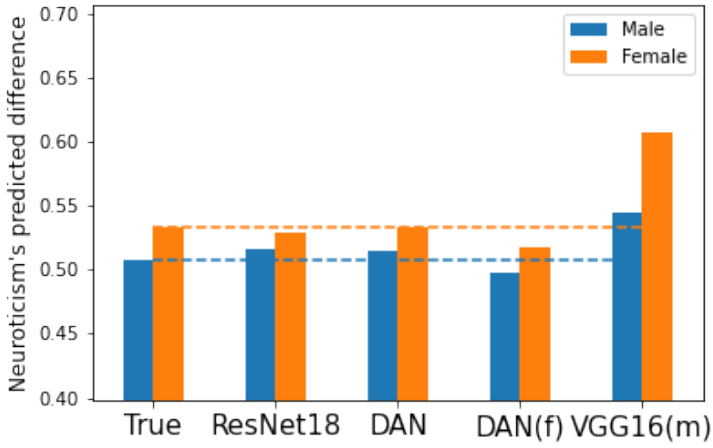


Figure 14. The mean predicted value for neuroticism for males and females compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

The accuracy of neuroticism for men is higher in the baseline, and the models that use the modified DAN structure, and for women, it is higher in the modified VGG-16 model. For neuroticism when looking at how the predictions compare (shown in Figure 14) to the actual value women are evaluated higher with the modified VGG-16 but with other models, males are comparatively rated as having higher neuroticism compared to the actual value.

6.2 Ethnicity

Comparing the predictions for different ethnic groups could be a little less accurate as opposed to the gender, where the number of both men and women is quite similar with slightly more women in the dataset. The ethnicity part of it is dominated by people of Caucasian descent, and people of Asian and African-American descent make up a minority of people in the dataset, but still, it is an avenue that can be looked at.

For different ethnicity groups, one thing that can immediately be seen is that the absolute accuracy for Asians is much higher compared to the other groups, which may come from the fact

that they make up a smaller minority in the dataset and every person in the test dataset

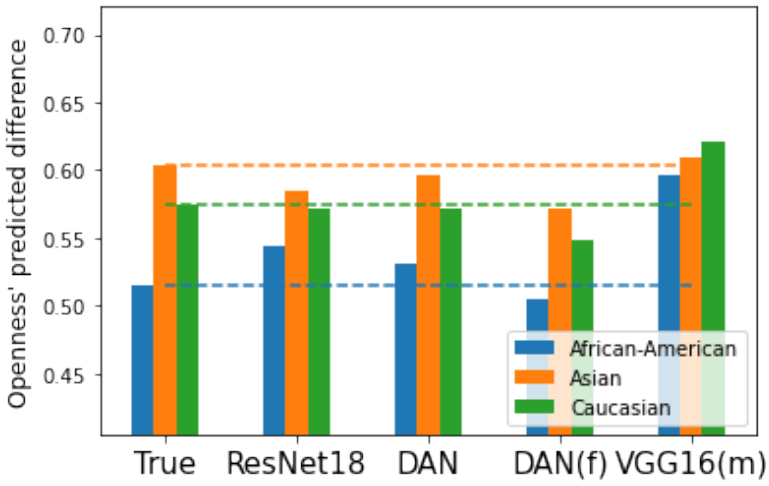


Figure 15. The mean predicted value for openness for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

is quite close to having average values for every trait, so there are not many people with very high or very low traits that may be predicted incorrectly. However, when comparing Caucasians and African-Americans to each other, then in some cases the overall absolute accuracy for Caucasians is higher and in other cases, it is higher for African-Americans.

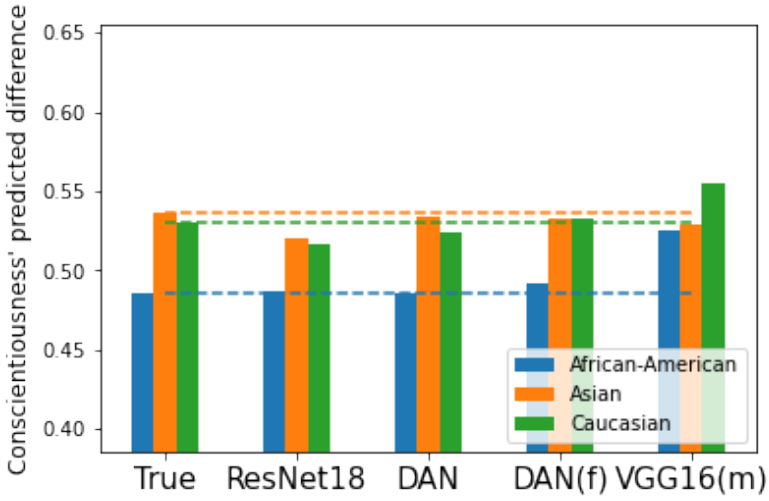


Figure 16. The mean predicted value for conscientiousness for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

Comparing the average predicted openness of each group to the actual values, shown in Figure

15, it is possible to see that compared to the actual value in reality Asians are predicted to have worse openness relative to the actual value and African-Americans seem to be predicted higher relative to the actual value. In all but the modified VGG-16 model, Asians are predicted to have lower openness than the actual value and even with the VGG-16 other groups are predicted to be relatively much higher when compared to Asians as on average here Caucasians are rated and the African-Americans relative to their actual average are also favoured. In the other predictions, it can be seen that relative to the actual value, Asians come out worst when considering how things should have been evaluated.

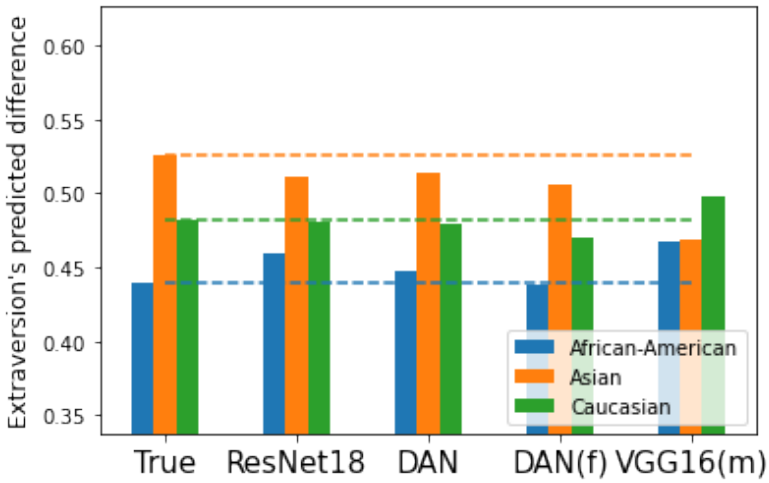


Figure 17. The mean predicted value for extraversion for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

Next when taking a look at how conscientiousness was predicted for each ethnic group then similar logic to openness can be seen here as well when looking at Figure 16, here it is visible that similar to the openness, Asians on average are seen by the models as having lower conscientiousness relatively compared to African-Americans and Caucasians, when taking into account the actual value.

In addition, to previously mentioned openness and conscientiousness when looking at the information shown in Figures 17, 18, and 19, it is possible to see that for extraversion, agreeableness and neuroticism, the relative prediction for each trait with Asians is comparatively much lower compared to Caucasians or African-Americans. Looking at the figures and concentrating on the African-Americans, then most often the prediction is higher compared to the average and in cases when it is lower than the actual value it is still comparatively higher than for Asians

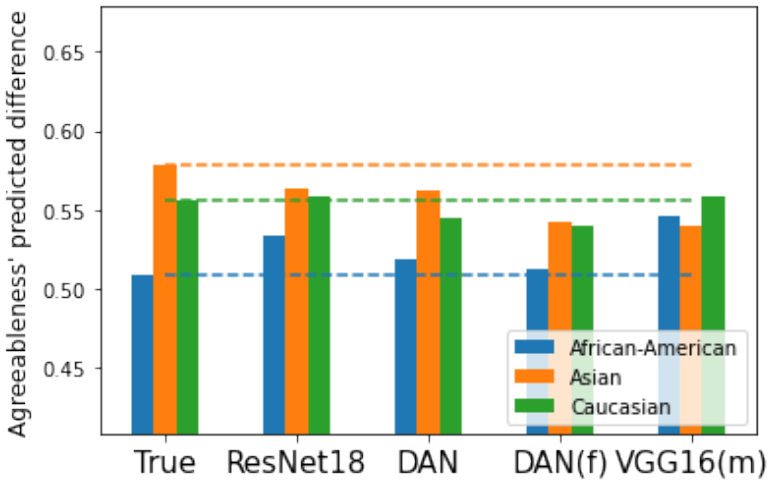


Figure 18. The mean predicted value for agreeableness for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

and Caucasians. In the Asian case, the values for each trait are comparatively lower than the actual value. The possible reason for the discrepancies in Asian and African-American data may be that they make up a minority in the dataset, as about $\frac{5}{6}$ of the videos in the training set is Caucasian and since their values are on average lower than Asians and higher than African-Americans, which could lead to lowering one group's prediction and increasing the other groups' prediction.

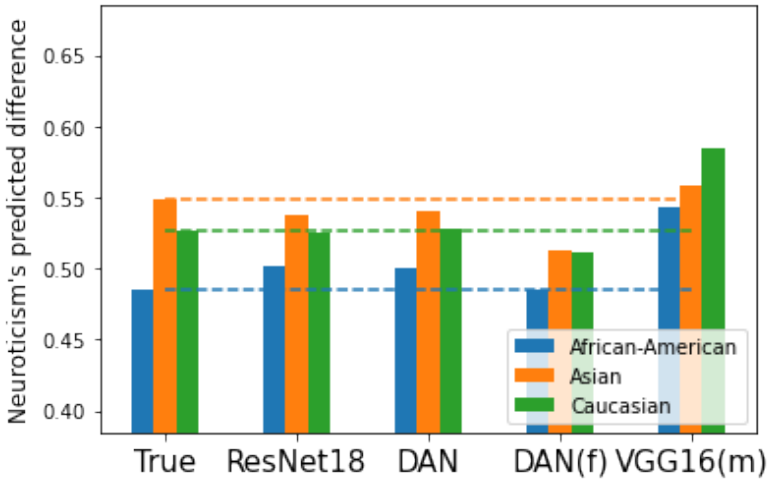


Figure 19. The mean predicted value for neuroticism for Asians, African-Americans and Caucasians compared to the ground truth. The dashed lines are to visualize the difference compared to the ground truth.

6.3 Balanced

6.3.1 Gender

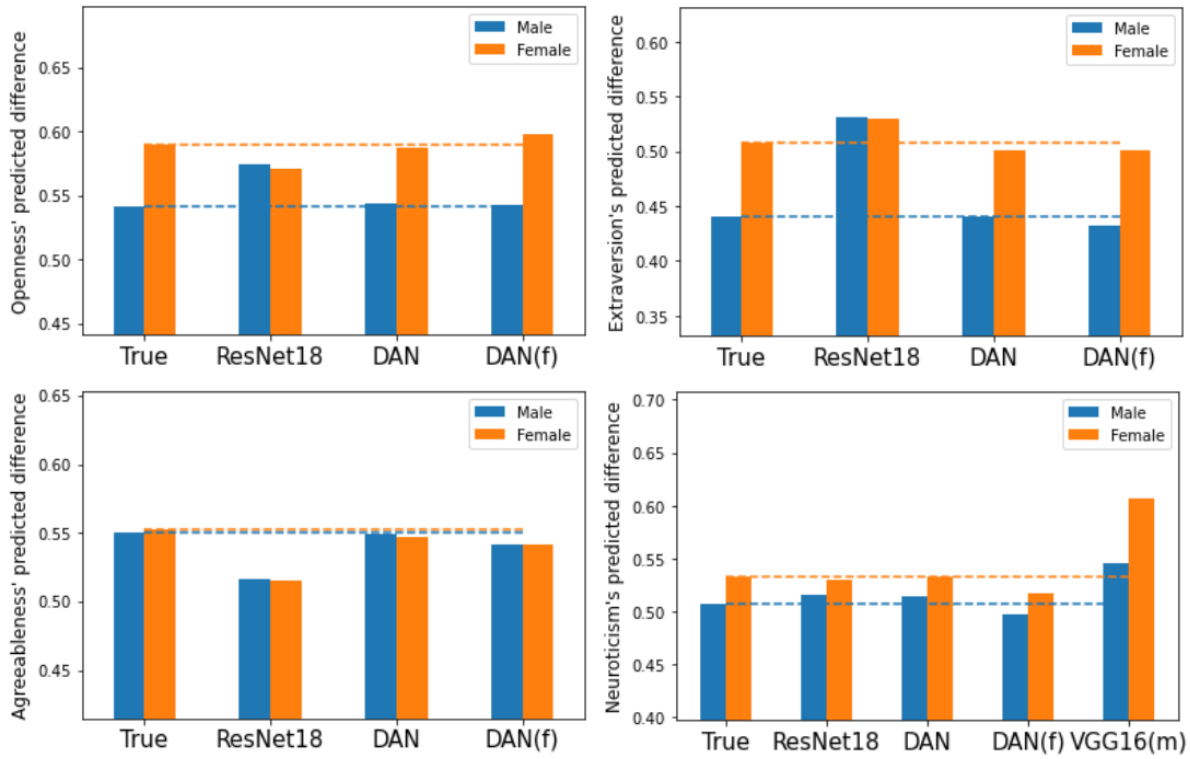


Figure 20. The mean predicted values for openness, extraversion, agreeableness and neuroticism compared to the ground truth for men and women in the balanced dataset.

Figure 20 shows the mean predicted values from methods where the balanced dataset was used for training when comparing it to previous methods where the datasets were imbalanced, then it is not possible to say if bias is more or less prevalent here. Conscientiousness is not shown with the balanced datasets, as the method using the facial region was not able to learn it and always evaluated it to be zero. The modified DAN methods worked similarly to the imbalanced version, and the ResNet-18 ended up worse, as both males and females were predicted very similarly to each other.

6.3.2 Ethnicity

Similarly to the means predicted for males and females from the balanced dataset, the mean values predicted for different ethnicities shown in Figure 21, it is possible to see that there is

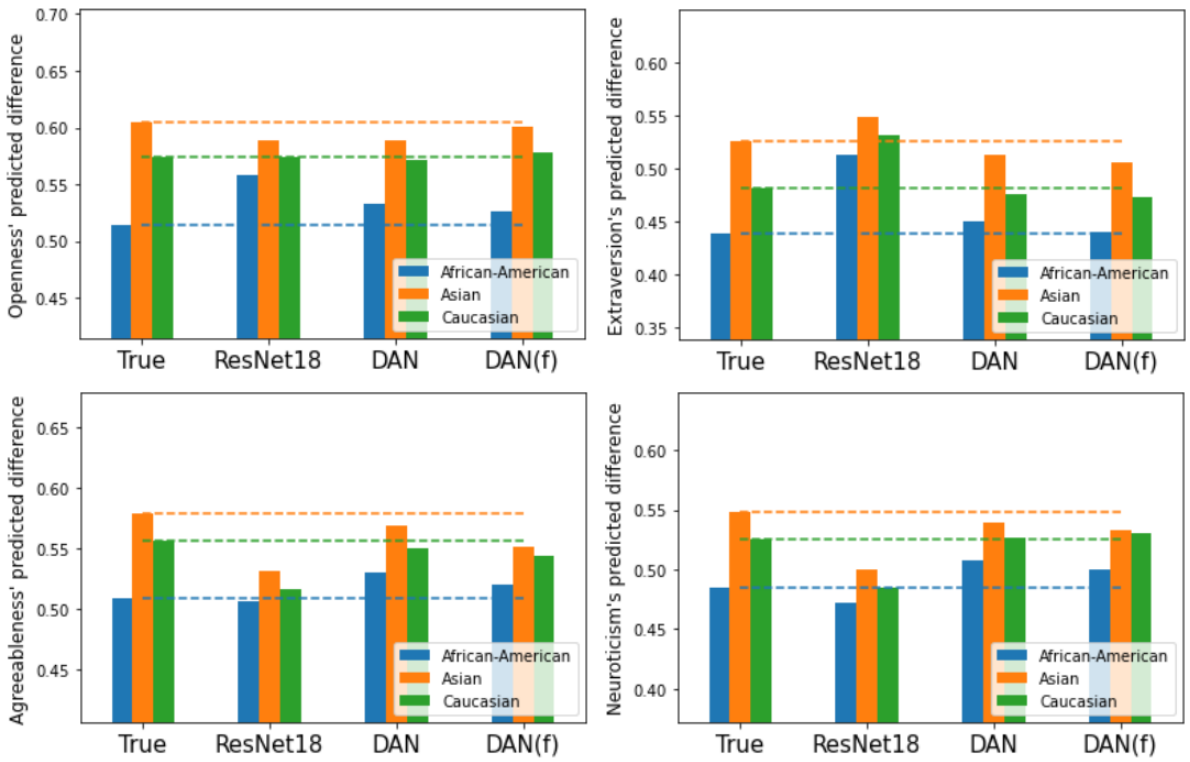


Figure 21. The mean predicted values for openness, extraversion, agreeableness and neuroticism compared to the ground truth for Asians, African-Americans, and Caucasians in the balanced dataset.

no discernable bias towards any group. Just like in the initial methods, the Asians have the highest absolute accuracy overall and the predicted mean is relatively still lower compared to African-Americans and Caucasians, so there is nothing definitive that can be concluded from these results.

6.4 Comparison

The average accuracy for all the models is brought out in Table 11. Firstly, when comparing all the methods for males and females the highest difference in accuracies between them was 0.003, so the models worked very similarly for both males and females. Secondly, looking at the accuracies for the three ethnicities represented in the dataset then in every situation the accuracy for Asians was the highest, however when comparing the accuracies in African-Americans and Caucasians then there is no certainty of whether one is more accurate than the other. The largest difference in accuracies between groups was 0.038. Compared to the gender

difference this is almost 13 times larger but, this mainly comes from the Asian group's higher accuracy, when comparing African-Americans and Caucasians their highest accuracy difference was 0.005, which is closer to the difference between males and females.

Table 11. Mean accuracies of different groups with different methods used.

Method	**Overall**	**Male**	**Female**	**Asian**	**African-American**	**Caucasian**
ResNet-18	0.90074	0.90134	0.90026	0.92194	0.89697	0.90063
Mod DAN	0.90605	0.90516	0.90675	0.92644	0.90179	0.90603
DAN face	0.91109	0.90942	0.91241	0.92664	0.91139	0.91061
Modified VGG-16	0.8633	0.86466	0.86222	0.89427	0.86548	0.86216
ResNet-18(b)	0.86988	0.86956	0.87015	0.90604	0.87266	0.86852
Mod DAN(b)	0.90478	0.90536	0.90431	0.92923	0.89974	0.90475
DAN face(b)	0.82081	0.82235	0.81957	0.8348	0.82385	0.82002

Looking at how it may have ended up that the highest accuracy was achieved with Asians, then when looking at the training dataset, then it shows that the variance of Asians is the lowest among all the different groups used in this thesis. Additionally, given that Asians comprised only 2.4% of the test dataset, then if there were not any major occurrences where the actual label was much higher or lower compared to the mean of the trait. Compared to the Asians, African-Americans had more occurrences of edge cases in the test set.

Summary

The area of interest in this master's thesis was to investigate the state-of-the-art deep learning-based methods in first impression analysis for ethnic and gender bias, which could result in unfavourable predictions for some groups. For this, a dataset containing videos of people speaking directly to a camera was used, these videos were given values to represent first impression personality traits using the big five personality traits. In addition, an additional table was used, which showed the gender, either male or female, and the ethnicity (Asian, African-American or Caucasian) of the person in the video. These were used to investigate possible biases between both males and females, as well as the three ethnicities.

Firstly, the dataset was used in a way that took the same number of frames from each video and those frames were used to train models to determine if some pattern existed about biases. This resulted in videos containing Asians being much more accurate on average compared to other ethnicities, but at the same time, their predicted mean values were comparatively lower than the actual value when compared to for example African-Americans, who on average had a mean predicted value higher than the true value. For the reason stated previously, the data used for training was changed, so that it would include similar amounts of every group, this data included similar amounts of males and females, which was already in the original data, but also the counts collected from the frames for each ethnicity would also be about equal to each other.

Secondly, the models were trained once more, but this time using input data so that every group would have equal representation. Models trained with this dataset did not fare any better compared to the initial models, and in reality, the results worsened.

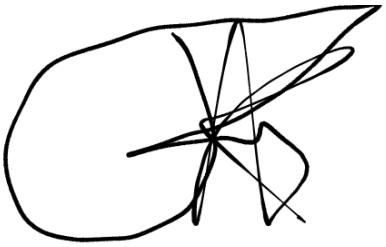
In conclusion, when comparing the accuracies achieved with different methods for different groups, then there was no significant difference between men and women. For ethnicity, the only visible diversion from the norm was with the Asian group, however, this may have arisen from the fact that there were not many examples of extreme values with the Asian group in the test dataset.

Future Work

This work only used visual information gained from the dataset, however, since the initial raw data is in the form of videos and since there exist methods capable of utilizing auditory and textual data as well, then this study can be extended to include such methods too. In addition, the dataset can be improved to include more ethnicities, as current data has limitations when looking at variety.

Acknowledgements

I like to thank my supervisors, Gholamreza Anbarjafari and Kadir Aktas. In Particular, Kadir Aktas, who guided me the whole way and was always ready to answer all of my questions and motivated me to get things done in a timely fashion. In addition, I would like to thank my family members, who have supported me wholeheartedly during this process.



Bibliography

[1] J. Hu, "Report: 99% of fortune 500 companies use applicant tracking systems," *Jobscan*, June 2021. https://www.jobscan.co/blog/99-percent-fortune-500-ats/#:~:text=Fortune%20500%20companies%20still%20use,of%20Fortune%20500%20market%20share., Accessed: 12.05.2022.

[2] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[3] Z. Zhang, M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda, H. Goyal, *et al.*, "Opening the black box of neural networks: methods for interpreting neural network models in clinical applications," *Annals of translational medicine*, vol. 6, no. 11, 2018.

[4] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in ai: the casual conversations dataset," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[5] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–16, 2019.

[6] T. Kasapoglu and A. Masso, "Attaining security through algorithms: Perspectives of refugees and data experts," in *Theorizing Criminality and Policing in the Digital Media Age*, Emerald Publishing Limited, 2021.

[7] R. D. P. Principi, C. Palmero, J. C. J. Junior, and S. Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 607–621, 2019.

[8] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. Jacques, M. Madadi, S. Ayache, E. Viegas, F. Gurpinar, A. S. Wicaksana, C. Liem, M. A. J. Van Gerven, and R. Van Lier, "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[9] A. Arya, L. N. Jefferies, J. T. Enns, and S. DiPaola, "Facial actions as visual cues for personality," *Computer Animation and Virtual Worlds*, vol. 17, no. 3-4, pp. 371–382, 2006.

[10] P. Costa and R. McCrae, "A five-factor theory of personality," *The Five-Factor Model of Personality: Theoretical Perspectives*, vol. 2, pp. 51–87, 01 1999.

[11] J. Willis and A. Todorov, "First impressions: Making up your mind after a 100-ms exposure to a face," *Psychological science*, vol. 17, no. 7, pp. 592–598, 2006.

[12] D. Schiller, J. B. Freeman, J. P. Mitchell, J. S. Uleman, and E. A. Phelps, "A neural mechanism of first impressions," *Nature neuroscience*, vol. 12, no. 4, pp. 508–514, 2009.

[13] J. C. S. J. Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. Van Gerven, R. Van Lier, *et al.*, "First impressions: A survey on vision-based apparent personality trait analysis," *IEEE Transactions on Affective Computing*, 2019.

[14] A. H. Sham, K. Aktas, D. Rizhinashvili, D. Kuklianov, F. Alisinanoglu, I. Ofodile, C. Ozcinar, and G. Anbarjafari, "Ethical ai in facial expression analysis: racial bias," *Signal, Image and Video Processing*, pp. 1–8, 2022.

[15] A. Das, A. Dantcheva, and F. Bremond, "Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach," in *Proceedings of the european conference on computer vision (eccv) workshops*, pp. 0–0, 2018.

[16] A. Vecvanags, K. Aktas, I. Pavlovs, E. Avots, J. Filipovs, A. Brauns, G. Done, D. Jakovels, and G. Anbarjafari, "Ungulate detection and species classification from camera trap images using retinanet and faster r-cnn," *Entropy*, vol. 24, no. 3, p. 353, 2022.

[17] D. Kamińska, K. Aktas, D. Rizhinashvili, D. Kuklyanov, A. H. Sham, S. Escalera, K. Nasrollahi, T. B. Moeslund, and G. Anbarjafari, "Two-stage recognition and beyond for compound facial emotion recognition," *Electronics*, vol. 10, no. 22, p. 2847, 2021.

[18] K. Aktas, M. Demirel, M. Moor, J. Olesk, C. Ozcinar, and G. Anbarjafari, "Spatiotemporal based table tennis stroke-type assessment," *Signal, Image and Video Processing*, vol. 15, no. 7, pp. 1593–1600, 2021.

[19] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[20] Y. LeCun, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation algorithm," in *Proceedings of NIPS*, 1990.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[22] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *2017 international conference on communication and signal processing (ICCSP)*, pp. 0588–0592, IEEE, 2017.

[23] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*, pp. 1–6, Ieee, 2017.

[24] L. Zaniolo and O. Marques, "On the use of variable stride in convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 13581–13598, 2020.

[25] H. Gholamalinezhad and H. Khosravi, "Pooling methods in deep neural networks, a review," *arXiv preprint arXiv:2009.07485*, 2020.

[26] F. Ramzan, M. U. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks," *Journal of Medical Systems*, vol. 44, 12 2019.

[27] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep bimodal regression for apparent personality analysis," in *European conference on computer vision*, pp. 311–324, Springer, 2016.

[28] C. Ventura, D. Masip, and A. Lapedriza, "Interpreting cnn models for apparent personality trait regression," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 55–63, 2017.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] D. Helm and M. Kampel, "Single-modal video analysis of personality traits using low-level visual features," in *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, IEEE, 2020.

[31] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions - dataset and results," *European Conference on Computer Vision*, 10 2016.

[32] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[33] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 59–66, 2018.

[34] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazare, T. Bui, and J. Collomosse, "Everything you wanted to know about deep learning for computer vision but were afraid to ask," in *2017 30th SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T)*, pp. 17–41, IEEE, 2017.

[35] S. Bhanja and A. Das, "Impact of data normalization on deep neural network for time series forecasting," *arXiv preprint arXiv:1812.05519*, 2018.

[36] "Percent difference," Cuemath, `https://www.cuemath.com/commercial-math/percent-difference/`, Accessed: 14.05.2022.

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Friedrich Krull,

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis
Assessment of ethnic and gender bias in automated first impression analysis,
supervised by Gholamreza Anbarjafari, and Kadir Aktas.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Friedrich Krull

17/05/2022