Ethical Implications of Face Recognition Tasks in Law Enforcement

Submitted in partial fulfillment for the degree of master of science

Jesse Tol
12112070

Master Information Studies
Information Systems

Faculty of Science
University of Amsterdam

July 19th, 2019

Universiteit van Amsterdam

1st supervisor
Dr. Sieuwert van Otterloo
ICT Institute

2nd supervisor
Dr. Frank Nack
Informatics Institute, UvA

ABSTRACT

This research investigates the effects of headgear on the performance of face recognition tasks by the Amazon Rekognition algorithm. Based on conditional requirements of fairness, and the definitions of predictive parity and predictive equality, the ethical implications of the Amazon Rekognition algorithm are examined. For this research, the Celebrity Headgear Dataset (CHD) is created, which enables the ability to measure the effects of headgear on the performance of face recognition tasks. The Amazon Rekognition algorithm is claimed as unfair as the performance results in processing the CHD does not achieve the requirements of fairness. The limited performance of the Amazon Rekognition algorithm on processing people with headgear implies discrimination due to a higher likelihood for the misidentification of people who are wearing headgear. Misidentifications in law enforcement could lead to the arrest of an innocent person, while the actual perpetrator remains free. Police departments should reconsider their usage of the Amazon Rekognition algorithm in law enforcement because unfair behavior could violate the individual's freedom and rights.

KEYWORDS

Artificial Intelligence, Computer Vision, Face Recognition, Ethics, Algorithmic Audit

1 INTRODUCTION

In the last decennia, artificial intelligence (AI) gained popularity in various domains. The AI Index Report by Shoham [19] measures the adoption since 1996 in several domains using multiple indicators. Between 2015 and 2018, the growth of active AI startups in the U.S. was 113%. In the same period, venture capital funding for AI startups in the U.S. reached a growth of 350%. AI also gained popularity in academia. Since 1996, the number of published papers about AI in government, corporate, and medical domains has been doubled seven times. The enrollment of students for introductory machine learning courses in the U.S. has been doubled five times between 2012 and 2017 [19]. The increased adoption of AI is caused by the introduction of innovative applications in various AI domains. Examples of AI applications are computer vision, speech recognition, and pattern recognition [3].

Within the computer vision domain, several intelligent algorithms have been developed by big tech companies [28], for example, IBM Watson [1], Microsoft Azure [2], and Amazon [3] have developed their computer vision algorithm. Each algorithm provides intelligent vision functionalities like face detection, face recognition, and facial expression recognition. These functionalities are publicly accessible via their APIs. Computer vision functionalities are used in several market segments.

Garvie [11] has published a report in 2016 about unregulated usage of face recognition applications in law enforcement in the United States (U.S.). More than 52 law enforcement agencies like the FBI, state, and local police departments, have confirmed their usage of face recognition systems to verify and identify the identities of suspected people to catch violent criminals. The application of face recognition in law enforcement tasks is beneficial. Due to the usage of face recognition, criminals and fugitives are arrested [11]. In 2016, at least 26 states in the U.S. had shared their database of profile pictures from driver's licenses and ID cards. Together, the law enforcement network contains profile pictures of more than 117 million American adults [11]. This network of images is used for the identification of suspected people. Due to this network, systems are developed that enable a police officer to identify a person on the street directly. With a photo from a smartphone, the system processes the picture directly, in order to verify the identity. In case of a verification, the system returns a response with the person's identity [11]. Such identifications are defined as Stop and Identify tasks, or, in case of an arrest at the police station, as Arrest and Identify tasks. The American Civil Liberties Union (ACLU) has revealed recently that the Washington county cooperates with Amazon to develop a mobile application that processes images instantly against a database of 300.000 mugshots [25]. Also, other police departments have confirmed they are using real-time surveillance video and image footage for the identification of pedestrians. If an individual or a group of individuals is suspected for a particular crime, the police uses face recognition tasks on multiple images and live video feeds to verify their presence. Based on these results, the police ascertain whether a suspected individual has crossed one of the locations that is covered by a surveillance camera. This type of face identification is called Investigation [11]. The Orlando Police Department applies the Amazon Rekognition algorithm for identifying faces real time using millions of faces [25].

1.1 Algorithmic Bias

However, face recognition algorithms are not supremely accurate and do not perform consistently [11]. Multiple factors are causing this lack of accuracy. The performance of face recognition is affected by changes in facial features due to age, cosmetics, glasses, and haircut. Poor quality images or differences in lightning is either affecting the performance of face recognition. A comparison of images with the same identity that is generated in different environments, like a mugshot in a controlled environment and a picture of a pedestrian from a surveillance camera, is also affecting the performance [11]. Besides the properties of pictures, multiple studies have concluded that face recognition algorithms are limited due to signs of racial and gender bias [11]. Face recognition algorithms perform poorly on African Americans, especially on females, in comparison with Caucasians [15]. The same defects occur for females in comparison with males, and for young adults in comparison with older adults [15]. The inconsistency in the performance of face recognition algorithms, implies unequal treatment among individuals dependent on gender and race. Six years after the study of Klare et al. [15], face recognition tasks have still problems with processing images of dark-skinned females in comparison with Caucasian's males. Recent research of Buolamwini et al. [4] in 2018, measures

[1] https://www.ibm.com/watson/services/visual-recognition/
[2] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/
[3] https://aws.amazon.com/rekognition/

the performance of three commercial gender classification algorithms on a dataset of 1270 parliamentarians from six countries balanced by gender and skin type. The three commercial gender classifications algorithms tested, were Microsoft's Cognitive Services Face API, IBM Watson's Visual Recognition API, and Face++'s facial analysis technology [4]. The results of this performance test are divided into four subgroups: light males, light females, dark males, dark females. This study states that the three algorithms performed worst on dark females, with an error rate between 20.8% and 34.7%. The accuracy was the highest on white skinned males, with error rates between 0.0% and 0.3% [4].

The application of inaccurate face recognition algorithms induces erroneous results in real-life situations. Racial and gender bias can imply discrimination due to unequal treatment among diverse groups. Regarding the usage of face recognition algorithms in law enforcement, an erroneous outcome causes the police to misidentify suspected people. A misidentification in law enforcement could imply an arrest of innocent people that have become qualified as suspected by the algorithm, while the perpetrator remains free. Due to the racial and gender bias that are ascertained in the report of Garvie [11] and in performance studies of Klare et al. [15] and Buolamwini et al. [4], this scenario is likely to occur when the suspected criminal is an African American female. The unequal treatment of individuals on the basis of race, religion, or any other ground, is called discrimination, which is not permitted according to the first article of the Dutch constitution [4].

1.2 Aim of this Research

Since the Amazon Rekognition algorithm is adopted in law enforcement, images of pedestrians with headgear regarding their religion or fashion, are processed for the identification of suspected criminals. In addition to current performance studies, which have only examined the performance of face recognition on race and gender, this paper aims to consider the effects of headgear on the performance of the Amazon Rekognition algorithm. Based on the performance regarding wearing headgear, the ethical implications of the actual usage of the Amazon Rekognition algorithm in law enforcement by US police departments are considered. For considering the ethical implications, this research provides three contributions. First, as specific face recognition datasets for measuring the performance of headgear are not available, this study introduces a new dataset, the Celebrity Headgear Dataset (CHD), which enables the ability to measure the effects of headgear on the performance of face recognition tasks. Second, for the first time, the effects of headgear on the performance of the Amazon Rekognition algorithm are measured specifically. At last, looking at the results of the CHD by the Amazon Rekognition algorithm, the ethical implications of the actual usage of face recognition algorithms in law enforcement are investigated. The ethical implications are considered based on definitions of fairness.

[4]http://www.wetboek-online.nl/wet/Grondwet/1.html

2 RELATED WORK

2.1 Artificial Intelligence

Artificial intelligence (AI) is a subject within computer science that focusses on systems that are intelligent by imitating the characteristics of human behavior. Artificial intelligence refers to intelligent systems that are able to perceive, reason and act based on configured principles[5]. Torresen [21] defines AI as a technology that can feel, reason and behave in the best possible way through learning. The goal of AI is to solve problems in society using advanced technology. These advanced technologies are tend to identify, understand, reuse, and surpass human intelligence [17]. AI researchers aim to define the elements of computations to identify intelligence and create machines that can behave intelligently. In a wider sense, AI tends to malinger the mental processes of human beings to provide a similar interaction between humans and computers [17].

The field of AI focusses on development methods like machine learning to enable innovative features. Such innovative features are the interpreting of spoken text and identifying objects in the real world [27]. A subset of machine learning is deep learning [12]. Deep learning intends to define input characteristics using representation learning. Based on the input characteristics, an output is composed. The process of defining input characteristics in order to generate output requires mathematical capabilities [12]. The mathematical processes aim to understand the relations of the input characteristics in order to generate customized output. Deep learning makes use of numerous interconnected neurons to facilitate mathematical capabilities. Every neuron executes mathematical processes. The output of the mathematical processes is a weight. The weights are similar to likelihood ratios in statistics [12]. Each neuron transfers the detected input values with the weight to the activation function. If the activation function surpasses the threshold, the input values including weight, become transferred to the other neurons [12]. If the threshold is not surpassed, the output will be zero [12]. Once the network provides an output, the network compares the output with the truth value that is defined in the label of the input object. The labels of the input objects are specified in the dataset. Erroneous weights are adjusted towards the label. The correction process is a mathematical function called loss function [12]. Through this process, the network intends to minimize the errors by updating the erroneous weights [12]. A DNN includes a massive amount of mathematical capabilities among multiple layers. Even the smallest DNN holds thousands of weights. The capabilities of DNNs enable the modeling of complex patterns in data [1]. The introduction of DNN enables the development of disruptive innovative features. One of them is computer vision, which allows machines to perceive, analyze, and understand the visual world. [12, 18, 27]. The Amazon Rekognition algorithm uses a deep learning framework to enable their computer vision functionalities [6].

However, despite the innovative possibilities of DNNs, the non-linear structure of DNNs induces lacks in transparency and explainability [18]. Due to a shortage of knowledge about the processes in an AI system, nobody can declare what data from the input image

[5]https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf

[6]https://aws.amazon.com/blogs/aws/amazon-rekognition-image-detection-and-recognition-powered-by-deep-learning/

derives the DNN to generate that type of decision [18, 25]. Therefore, the DNNs are called black boxes [1, 18, 25]. The approach to experiment with innovative systems on real populations, without knowing the decisions of the DNN, and to investigate the implications afterward, limits the development of AI [25]. An example of a limitation is an erroneous decision of the DNN in self-driving cars, which could have dangerous consequences for the safety of pedestrians.

In order to guarantee the capabilities of AI systems that use DNNs, explainable AI is required [18]. Explainable AI aims for transparency of the processes within AI systems that declare the motives behind their reasoning. The explanation of their reasoning enables understanding and verification of the decisions made. Based on declarative reasoning of the AI system, flaws like biases could be detected and eliminated [18].

2.1.1 Computer Vision. Since the 1990s, the research and usage of computer vision have increased [28]. Due to the introduction of new technology and media, new applications and methods are originated that has evolved human-computer interaction. The decrease in the price-performance ratio of computing has implied the development and adoption of computer vision in current systems [26]. Face recognition technology evolves human-computer interaction due to its ability to communicate and behave naturally in human interaction [16].

Computer vision contains multiple functionalities like face detection, face recognition, and facial expression recognition. The main goal of face processing is that the computational behavior acts based on the information about the human's identity and intention that is extracted from the image [28]. Facial feature detection identifies facial elements like eyes, nose, mouth, and lips in an image. Face recognition intends to recognize the face on an input image in a database of images. Moreover, face authentication aims to verify the identity of an individual among the input image [28]. DNN empowers these computer vision functionalities. The DNN includes numerous neurons that provide a huge amount of mathematical capabilities for understanding the characteristics of the input image [12].

The first step within face processing is face detection that tends to detect a human's face in an image. The goal of face detection is to scan human faces and provide the location and extent of the detected faces [26]. Face detection is one of the most commonly known topics within computer vision due to various applications that have become enabled by face detection [26]. In addition to face recognition, facial expression recognition identifies the type of emotions in the detected faces. In the process of recognizing facial expressions, a separation is made between the signal, which refers to the facial movement, and the message, which refers to the meaning of the signal [16]. Implicit tagging is based on facial expression recognition and categorizes the type of multimedia content using tags by identifying the reaction of the watchers using facial expression recognition [16].

Yang et al. [26] provide four methods of single image detection: knowledge-based methods, feature invariant approaches, template matching methods, and appearance-based methods. Knowledge-based methods are rule-based systems that encode human knowledge for defining the unique characteristics of a face. These rules are derived from the researcher's knowledge to establish the connection between facial features for face localization [26]. Feature invariant approaches intend to first detect standard facial features like eyebrows, eyes, nose, and mouth, to consider the presence of a face. Template matching methods use standard face patterns to identify the face as a whole. Face detection and localization are based on the similarities between the input image and the implemented face patterns [26]. Appearance-based methods are mainly used for face detection [26]. In contradiction to template matching methods, the conditions of facial appearance are not predefined by experts but are considered based on a set of training images. Appearance-based methods rely on statistical analysis and machine learning methods to determine the conditions of images that contain faces. Then, based on these conditions, faces are detected in input images [26].

Face processing in images is challenging due to differences between images. Elements like pose, facial expression, occlusion, and image conditions can differ among images, which makes it hard to detect faces correctly [26]. Hassaballah et al. [14] classify five challenging categories. The first category is illumination variations, which refer to factors like lightning and camera capabilities that may affect the appearance of a human face in an image. Illumination variation implies the problem that the face of a person is presented significantly different. The second challenging category refers to the various poses of a person among several images. Different poses can affect the visibility of facial features like eyes and nose, which can lead to projective deformations and self-occlusion. The third category is related to aging, which can induce facial differences when a person becomes older. Facial recognition is challenging when the unique facial characteristics of an old image are changed in real life through aging. The fourth category refers to facial expressions. The recognizability of faces is dependent on the person's facial expression; changes in hair such as hairstyle, beard or mustache, affect the identifiability of a person. The fifth category is occlusion. Other objects can partially occlude facial elements. Images that present multiple people, faces, or other objects can hide certain elements which affect the identifiability of human faces [14].

2.2 Face Recognition Analysis Benchmark

The National Institute of Standards and Technology (NIST) is a non-regulatory federal agency, and part of the U.S. Department of Commerce. The NIST aims to stimulate innovation that fosters economic security and quality of life, by providing extensive measurements in science, standards, and technology. One of the projects hosted by NIST is the Facial Recognition Vendor Test (FRVT). The FRVT measures the accuracy and speed of automated facial recognition technologies that are used in real-world use cases like law enforcement and homeland security applications [13]. NIST has executed two types of reports: FRVT for verification (1:1 one-to-one) and identification (1:N one-to-many). The FVRT is initiated for developers, end users, and others who are interested in face recognition technology. Developers can submit their technology to NIST for a performance assessment. Participation does not charge submission fees and is open worldwide. The FRVT 1:N identification report investigates the performance of identification tasks. Identification tasks aim to detect the identity of an individual among a

database of numerous images. In identification tasks, a single input image is used to define the individual's identity, in order to check whether this identity occurs on the enrolled images [13]. The latest FRVT 1:N Identification report was published in November 2018. This report presents the performance results of 127 algorithms submitted by 45 developers. The performance test uses four datasets with different types of images. In total, the performance is measured based on 30.2 million copies from 14.4 million individuals. The input images are mugshots, poor quality webcam images, and images from surveillance videos [13]. The latest FRVT 1:N Identification shows, in comparison with the FRVT 1:N Identification report from 2013, a substantial decrease of the false negative identification rate (FNIR). 95% of the searches that failed in 2013 by providing the wrong person at rank 1 showed the correct result in 2018. The FNIR for the Microsoft-4 algorithm decreased from 4.1% in 2013 to 0.23% in 2018. For the same Microsoft-4 algorithm, the FNIR was 15.6% on mated-searches with a threshold in 2018. The increase of the FNIR when using a threshold is caused by poor image quality, aging, and presence of lookalikes [13].

The FRVT 1:N report separates two types of errors: type 1 and type 2 errors. A type 1 error emerges when the identity of the input image is matched with one or more different identities, so-called false positives. A type 2 error, also called as a miss, emerges when a search of a mated input image, is not matched with the identity of the input image. The metrics for measuring the type 1 and 2 errors are the False Positive Identification Rate (FPIR) and the True Positive Identification Rate (TPIR). A mated-search can return multiple non-mated candidates above the threshold. In addition to the FPIR, the Selectivity (SEL) quantifies the average number of non-mated returns for mated-searches. SEL divides the total number of non-mated candidates of mated searches, by the number of mated searches [13].

3 ACCURACY, FAIRNESS, AND BIAS

3.1 Bias

The adoption of AI algorithms in real-world applications is rising. Nowadays, algorithms enable automated decision making in various forms. AI is replacing humans in decision making processes like accepting loan requests and deciding which applicant gets hired for a specific job [20, 23]. Autonomous systems can decide without human intervention. The extension of autonomous systems in real-world use cases increases the necessity for verifying their outcomes to ensure decent behavior [21].

Algorithms use historical datasets in order to train themselves for composing a decision making process [2]. In Deep Neural Networks (DNN) the output is agreed on the labels of a dataset to correct erroneous weights [12]. Therefore, it depends on the integrality and quality of the dataset, whether the DNN is updating the weights on correct labels [2]. If the training datasets of DNN include subjective data from humans, DNN is trained on incomplete data. Incorrect labels could affect the DNN to involve wrong labels into the model, which, in the end, could produce erroneous decisions [2]. The usage of algorithms in various domains has shown flaws due to the existence of racial and gender bias. Bias is either originated through biased configuration during development or by the utilization of biased datasets [20]. A biased example regarding speech recognition was an algorithm that was underperforming on dialects since it was not extensively trained on dialects. Bias implies defects in the performance of the algorithm because it could generate and adapt faulty decisions that affect the subjects negatively [7].

3.2 Fairness

Similar to the adoption of AI algorithms, is the increasing attention of AI ethics. AI ethics strives for avoiding adverse effects by implementing the occasion for human inventory and the ability of self-reflection by the system themselves based on ethical perspectives [21]. The Platform voor InformatieSamenleving (ECP)[7] qualifies an AI application as ethical when the application maintains the well-being of people and planet. ECP has published the Artificial Intelligence Impact Assessment (AIIA) to support companies in evaluating the ethical status of their applications [8]. The AIIA maintains ethical conditions like integrity, safety, and transparency. The AIIA includes eight questionnaires that cover the urgency, explainability, consequences, trustworthy, and transparency of AI usage. Through these eight steps, the ECP intends to arouse the initiators for considering the ethical effects of their applications. One of the ethical principles of the AIIA is fairness.

Algorithmic fairness strives for preventing people from being mistreated based on personal characteristics [8, 20]. Algorithmic fairness is introduced to assess the behavior and effects of the algorithm's results. This research defines fairness following the definition of the Oxford Advanced American dictionary [9]: "Fairness is the quality of treating people equally, or in a way that is reasonable." Article 21 of the General Data Protection Regulation (GDPR) affirms the right of the subject to not be part of a decision that legally affects the subject, and which is created by solely automated processing [10]. Unfair algorithms can imply discriminative outcomes. Discriminative decisions disadvantage an individual or a group of people based on race, gender, religion, or on any other aspect. If headgear affects the performance of the Amazon Rekognition algorithm, the algorithm could disadvantage religious people who are wearing headgear according to their religion. Dolam [8] separates direct and indirect discrimination. Direct discrimination disadvantages an individual, in comparison with others. Indirect discrimination enforces the disadvantage of people due to a decision that is made previously [8].

Discrimination is not permitted globally. The first article of the Dutch constitution states for equal treatment of people. Discrimination on the basis of gender, race, religion, or any other ground is not permitted in the Netherlands [11]. The anti-discrimination law is also implemented in article 21 of the EU Charter of Fundamental Rights [20].

3.2.1 Measuring Algorithmic Fairness. Regarding the usage of face recognition tasks within law enforcement, the definition of fairness requires similar treatment of similar people, with reasonable effects

[7] https://ecp.nl/

[8] https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf

[9] https://www.oxfordlearnersdictionaries.com/definition/american_english/fairness

[10] http://www.privacy-regulation.eu/en/article-22-automated-individual-decision-making-including-profiling-GDPR.htm

[11] http://www.wetboek-online.nl/wet/Grondwet/1.html

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Figure 1: Confusion Matrix

for the affected subjects [6]. In order to maintain the reasonability of the impact in law enforcement, conditional requirements of fairness are set. The conditional requirements are based on the performance of a similar identification method in law enforcement. For more than 100 years, forensic fingerprints are used for criminal identification within law enforcement. In fingerprint examination, latent prints are validated against a database of other latents to identify the same DNA [22]. The performance study of Ulery et al. [22] measured the performance of examiner conclusion regarding forensic fingerprints. The performance study sets a false positive error rate of 0,1%, and a positive predictive value of 99,80%. Since forensic fingerprint is applied for a long time as an identification method in law enforcement, this research requires a minimal equal performance for face recognition tasks within law enforcement.

According to the definitions of fairness of Verma et al. [23], an algorithm is measurable on fairness. An algorithm behaves fairly when their performance results are satisfying either group fairness, predictive parity, or predictive equality, dependent on the type of face recognition task. A confusion matrix enables the possibility to consider the fairness of an algorithm quantitatively [20, 23]. Group fairness endeavors a similar ratio between subgroups of being predicted to the correct class. Group fairness requires an equal probability for every individual being predicted correctly, regardless of the individual's characteristics. Group fairness requires a similar acceptance rate and is based exclusively on predictive values [20, 23]. In addition to group fairness, predictive parity is looking at both the predicted positives and negatives. Predictive parity investigates the number of results that are predicted wrongly. Predictive parity requires an equal positive predictive value (PPV) among the different subgroups in the dataset. An equal PPV means that probability of being predicted correctly based on the actual value is similar for every group in the dataset [20, 23]. An extension of the predictive parity is predictive equality. Predictive equality requires the same false positive rates (FPR) for all subgroups. The predictive equality points for an equal proportion between the subgroups of actual negatives that are predicted positive [20, 23]. When a classifier does not accomplish the requirements of any metric that is relevant for a particular use case, it is claimed to be unfair. By investigating the performance results using a confusion matrix, AI algorithms are assessed on fairness. The definitions of fairness observe the occurrence of biased behavior [23].

3.3 Accuracy

A confusion matrix is common for measuring the accuracy of classification models. The confusion matrix divides the results into true and false results. Figure 1 presents the confusion matrix. True and false results are considered based on the checking whether actual negatives and positives are also predicted as negatives and positives [23, 24]. The True Negatives (TN) are the results that are annotated the same as how the algorithm has predicted them. Regarding the face recognition performance test, the actual negatives are the non-mated images that do not belong to the image collection of the input image. A result is a true negative (TN) if a non-mated candidate is predicted correctly as non-mated. When a non-mated image is returned as a mated image, the candidate is classified as a false positive (FP). A candidate is categorized as a false negative (FN) when a mated image is returned by the algorithm as a non-mated image. At last, true positives are the mated images that are processed as mated images [24].

The calculation of the algorithm's accuracy uses the confusion matrix. The accuracy metrics are: Positive Predictive Value (PPV), False Discovery Rate (FDR), True Positive Rate (TPR), and False Positive Rate (FPR). The PPV is the fraction of the true positives that are predicted the same as the total actual positives. The PPV indicates the precision of the algorithm [23, 24]. The formula of the PPV is:

$$PPV = (TP + TN)/TotalEnrolled \tag{1}$$

The FDR is the opposite of the PPV, which calculates the proportion of the incorrect predicted false positives over all predicted positives [23]. The FDR indicates how often the algorithm classifies the search incorrectly. The formula for calculating the FDR is:

$$FDR = (FP + FN)/TotalEnrolled \tag{2}$$

The TPR measures the proportion of actual positives that are returned by the algorithm as positive [23, 24]. The formula of TPR is:

$$TPR = TP/(TP + FN) \tag{3}$$

The FPR measures the proportion of actual negatives that are returned by the algorithm as positive [23, 24]. The formula of FPR is:

$$FPR = FP/(FP + TN) \tag{4}$$

4 USAGE IN LAW ENFORCEMENT

According to the NIST, face recognition tasks in law enforcement can be divided into three types of use cases: Investigation, Negative Identification, and Positive Identification [13]:

- **Investigation**: Examines the presence of a person on a single input image, in an authoritative set of images. Beforehand, it is unclear if, and how many times, the person on the input image is present in the set of images. The algorithm provides a configured number of potential candidates who are likely to be the input image's person. In most cases, the results get screened by a human to determine whether which results are correct matches. Based on correct matches, the identity of the input image is ascertained. This type of use case is characterized by small search volumes like one input image, and the application of human intervention for considering matches. The investigation is used within law enforcement processes to identify the criminal's identity based on a single mugshot.
- **Negative Identification**: Negative identification is the type of use case that assumes that the person on a given input image is not present in a set of images. This use case is characterized by high search volumes and limited time for human intervention. Therefore, a threshold is applied so that

the recognition algorithm only provides candidates who are very likely to be the same person in comparison with the input image. If the recognition algorithm does not provide any candidates, because all similarity scores are below the threshold, it is assumed that input image's identity does not occur in the dataset.

- **Positive Identification**: Access control is a typical application of a positive identification; the person's face gets access if the face matches with any enrolled identity. The face recognition algorithm uses a threshold to exclude false positives. When the person is not an enrolled identity, it is assumed that the similarity score will not cross the threshold.

The accuracy in investigation usage is measured based on rank-based metrics. For negative and positive identification, accuracy is calculated based on threshold-based metrics.

5 FACE RECOGNITION DATASETS

This section describes the two datasets that are used within this performance test. Since the NIST has executed several performance tests of face recognition algorithm using images similar to the The Multiple Encounter Dataset (MEDS-II), the MEDS-II functions as benchmark in this research. Based on the limitations of the MEDS-II, within this research, the Celebrity Headgear Dataset (CHD) is composed.

5.1 Multiple Encounter Dataset

The National Institute of Standards and Technology (NIST) has published the NIST Special Database 32 to execute their NIST Multiple Biometric Evaluation and to promote other researches [10]. The NIST Special Database 32 MEDS is originally published in 2010 as MEDS-I. In 2011, an updated version, called MEDS-II, is published that has doubled the number of images and extended metadata to support research in pose definition and local face features. The dataset has been used by the FBI and partner organizations to develop, test, and optimize face recognition applications. MEDS-II is suitable for various judicial applications like forensic comparison and face image conformance [10].

The MEDS-II contains 1309 images of 518 unique subjects. The images are extracted from 518 deceased persons. The number of encounters and time interval between multiple encounters varies per individual [10]. The labeling methodology of the MEDS-II is executed by NIST. Race and gender labels are defined based on observations or provided by the subjects themselves. Gender labels 'M' and 'F' are used to define males and females respectively. Race is defined as Asian ('A'), Black ('B'), American Indian ('AI'), Unknown ('U'), or White ('W'). The distribution of race and gender is presented in figure 2 [10]. From the 518 subjects in MEDS-II, 119 subjects are selected for the facial recognition audit. The 119 input images include both mated and non-mated search. To realize an audit result that is proportionally distributed, the same number of males and females are picked for both races. Figure 5 presents a specified overview of the 119 individuals from MEDS-II.

The images from the MEDS-II are mugshots and webcam images. Mugshots are characterized by constrained portraited-styled profile pictures with contrasting backgrounds, and a clear presentation of the person's face. Mugshots are frontal images on where the person is looking in camera direction. In figure 3, six mated mugshots are presented, that are part of the MEDS-II.

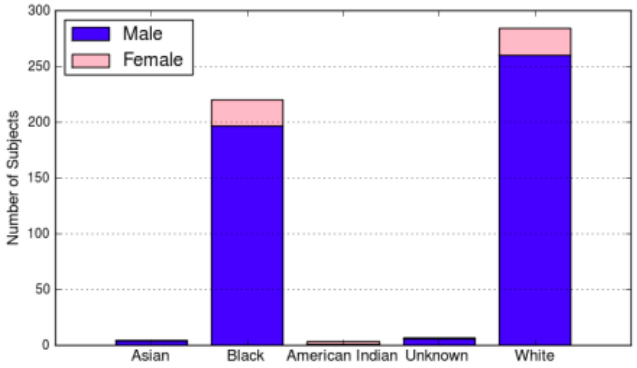


Figure 2: Distribution of Race and Gender in MEDS-II [10]

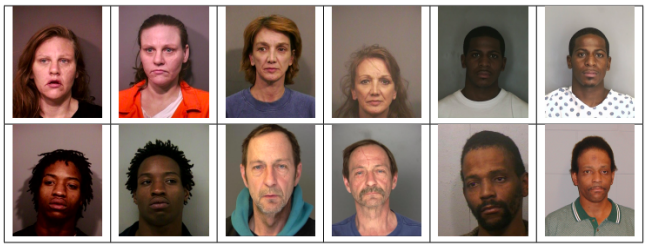


Figure 3: Six Mated Mugshots from MEDS-II [13]

Figure 4: Twelve Webcam Images from MEDS-II [13]

5.1.1 Limitations of the MEDS-II. The MEDS-II does not include any headgear. However, as face recognition tasks are applied in law enforcement using surveillance camera footage to identify pedestrians, it is likely to occur that pedestrians with headscarf, hat, or cap are processed by the algorithms. Therefore, it is crucial to test and maintain the performance of face recognition tasks on headgear. Erroneous identifications caused by headgear could lead to arrests of innocent people due to misidentifications, while the perpetrator remains free. As MEDS-II does not capture the people with these garments, within this study, the Celebrity Headgear Dataset (CHD) is composed.

5.2 Celebrity Headgear Dataset

This study introduces a new dataset based on the limited availability of face recognition datasets that cover the appearance of headgear. Since headgear is part of popular religions like Islam and Judaism [5], and headgear is involved in fashion, there is a necessity

MEDS-II	Black		White		Total
	M	F	M	F	
Mated	25	11	23	12	**71**
Non-Mated	12	12	13	11	**48**
Total	**37**	**23**	**36**	**23**	**119**
Headscarf	0	0	0	0	**0**
Cap	0	0	0	0	**0**
Hat	0	0	0	0	**0**

Figure 5: Overview of the Images from MEDS-II

CHD	Black		White		Total
	M	F	M	F	
Mated	21	20	40	20	101
Non-Mated	0	0	0	0	0
Total	**21**	**20**	**40**	**20**	**101**
Headscarf	10	19	25	19	73
Cap	4	0	6	0	10
Hat	7	1	9	1	18

Figure 6: Overview of the Images from CHD

for testing and maintaining the performance of face recognition tasks on headgear. Due to the limited availability of face recognition datasets that cover these requirements regarding headgear, this study introduces a new dataset called the Celebrity Headgear Dataset (CHD).

The difficulty of face recognition datasets is the requirement to gather multiple images of the same individual. Regarding face recognition usage, it is needed to adopt challenging images by varying the image's conditions. To solve this difficulty and fulfill the requirements, the CHD uses an inventive approach by picking multiple images from celebrities like actors, fashion models, or professional athletes. The beneficial property of actors is the availability of numerous images on the Web and the various attributes from the catwalk and movie scenes. The CHD makes use of the availability of images from the same actor with multiple properties in different circumstances. As the individuals are disguised for fulfilling the movie character, the CHD does not include sensitive information of ethnical minorities. Since the CHD's images are gathered from media productions and publicly accessible, the usage of these images does also not affect their privacy.

The CHD includes 628 images of 101 individuals. The image collection of every individual contains at least one image that presents a type of headgear. The type of headgear is defined for each image as either headscarf, cap, or head. The CHD also defines gender and skin type as black or white. White males are the majority in the CHD with a 40% share. The white females, dark males, and dark females are fulfilling the remaining 60% with each an equal proportion of 20%. On average, every image collection holds 6.2 images. Figure 6 presents an overview of the input images of the CHD.

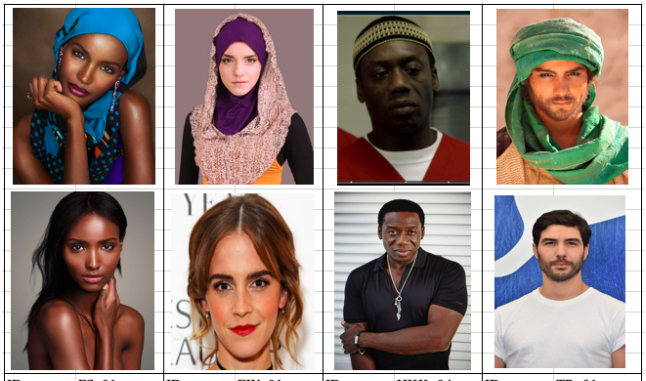
5.2.1 Gathering Images. The 101 celebrities are either actors, fashion models, or professional athletes. The benefits of these types of professions are the availability of various images on the Web. The image collection is composed using images of movie scenes, catwalks, sports matches, and model pictures. These types of images provide profile images under different circumstances and are publicly available. IMDB [12] is an online database that provides data about TV, movies, and celebrities. For famous actors, IMDB provides a profile page, including a photo gallery. Images from Arabic actors with and without headscarf are gathered from IMDB, using a list of Arabic actors[13]. All images from models are gathered via the search function of Google [14]. Various terms are applied in Google search, like "African models with a headscarf" or "African American baseball players". Google provides related images for the images that are selected. Via this related search function of Google, images of other actors, fashion models, and professional athletes are found. All images are downloaded and saved exclusively in a directory on Google Drive [15]. For every image collection, two noise images are added to make the dataset more challenging for face recognition tasks. Noise images are gathered using Google's upload image function in Google search. This search function provides a list of "very similar images". For personal preferences, the Google Chrome browser [16] is used all the time during the gathering images process of the CHD. The complete dataset is accessible on Google Drive [17].

5.2.2 Labeling Methodology. The metadata document specifies gender, skin type, and type of headgear. All data is manually annotated by the creator themselve. The gender of every individual is determined based on the biographical information provided by the Google search results. Males and females are annotated by the labels 'M' and 'F' respectively. Transgenderism is not included in the gender labeling.

Besides the type of headgear and gender, the CHD also defines the skin color of every individual. Inspired by the skin type labeling methodology of Buolamwini et al. [4], the criteria of black and white skin types are based on the Fitzpatrick six-point labeling system [9]. Fitzpatrick [9] published a framework to classify skin types based on sun reaction. Six skin types (I-VI) are defined. The whitest skin type, formulated as I, gets easily sunburned and does not becomes tan. The darkest skin type, formulated as VI, is black and never gets sunburned. To annotate every individual as either black or white, CHD defines a white skin (W) based on Fitzpatrick's skin type I, II, III, and defines a black skin (B) based on Fitzpatrick's skin type IV, V, VI. Due to this definition, the difference between the two skin types is the possibility for white skins to become sunburned after sun exposure [9].

Three types of headgear are defined within the CHD. A headscarf is covering mostly the top and sides of a person's head using a cloth. A headscarf is characterized by a cloth that is wrapped around the individual's face. A cap is only covering the top of a person's head and is wearing tight. A cap is characterized by the protrusion on the

[12] https://www.imdb.com/
[13] https://www.imdb.com/list/ls045413236/
[14] https://www.google.com/
[15] https://www.google.com/drive/
[16] https://www.google.com/chrome/
[17] https://drive.google.com/open?id=1cZLnWdGZcCpwo9c2I4ym07z7T95cYvAj



ID:	FS_01	ID:	EW_01	ID:	HKK_04	ID:	TR_01
Subgroup:	Dark Female	Subgroup:	White Female	Subgroup:	Dark Male	Subgroup:	White Male
Headgear:	Headscarf	Headgear:	Headscarf	Headgear:	Hat	Headgear:	Headscarf
Similarity:	97,86%	Similarity:	86,86%	Similarity:	97,40%	Similarity:	89,07%

Figure 7: Example Images of CHD Including Similarity Scores

ID:	AD_01	ID:	AM_01	ID:	AVL_01	ID:	ZS_01
Subgroup:	Dark Female	Subgroup:	White Male	Subgroup:	White Female	Subgroup:	White Male
Headgear:	Headscarf	Headgear:	Headscarf	Headgear:	Headscarf	Headgear:	Hat
Similarity:	14,70%	Similarity:	20,98%	Similarity:	22,86%	Similarity:	46,74%

Figure 8: Mated Images of CHD with Lowest Similarity Scores

front side that aims to cover face for sunlight. CHD has included images from baseball players with a cap to measure the effect of caps in face recognition tasks. At last, hats are annotated in the CHD. Hats are similar to caps that only cover the top of the person's head, but without the protrusion to cover the sun. The headgear on the input images determines type of headgear on an individual.

Figure 7 shows an example of mated searches of the CHD. The details below the images specify the name, ID, subgroup, and the type of headgear for every individual according to the labeling methodology. The last column presents the similarity score generated by the Amazon Rekognition algorithm.

Figure 8 presents mated searches of the CHD with the lowest similarity scores. Since all four similarity scores are below the threshold of 55,92%, all were categorized as false negatives.

5.2.3 Corrections. During creation of the CHD, we discovered that the labeling methodology can make mistakes. One of the corrections contained to the image collections of AEM and AES. The image collections of AEM and AES both contained a mated image that presented more than one person. In case of multiple people on a picture, the head with the biggest size is picked for the face recognition task. This determination is configured in the python program. For the image of AEM and AES, the Amazon Face Rekognition algorithm picked the wrong face, which implied a low similarity score for a mated image. These images are edited by cutting off the other face(s) to ensure the right face is picked for the face recognition task. After editing, the similarity score between AES_04 and AES_02_input increased from 14,49% to 99,10%. The similarity score between AEM_02 and AEM_03_Input increased from 19,49% to 97,70%. The image collection of CE included two noise image that both presented the identity of DO. Both noise images of CE are changed into additional mated images of DO. The same situation occurred for one noise image of AD, which is changed into an extra mated image of FS. The corrected annotations are:

- CE_06_Noise = DO_08
- CE_07_Noise = DO_09
- AD_05_Noise = FS_08

After reconsidering the mated searches, an erroneous mate was found in the image collection of KN. One of KN's mated images presented a different person, which caused a low similarity score for this mated image. This image is changed into an extra noise image of KN: KN_04 = KN_04_Noise.

The manual process for annotating the CHD is a time-intensive and error-prone effort. For extending the CHD, the labeling methodology could be executed automatically. For example, gender classification algorithms could be used for annotating gender. After annotating the whole dataset, examination for errors is required to prevent usage of incorrect labels in the dataset.

5.2.4 Metadata. The CHD includes a metadata document that specifies the celebrity's name, CHD ID, Number of matches, number of noise gender, skin type, and type of headgear. The ID is created based on the letters of the first and last name. For each individual, a directory is created to separate the image collections. The structure of every image is ID_NR. The input images are defined by _input after NR. Noise images are defined by _Noise after NR. Figure 6 specifies the input images of CHD.

5.3 Overview Datasets

Figure 9 presents an overview of all images of MEDS-II and CHD that are part of the audit. In this table, the differences in numbers between the two datasets are visible. Both datasets include the same proportion of males (61%) and females (39%), In total, approximately the same proportion of black and white people are applied in the performance test. The input images are tested among the enrolled images. In MEDS-II, to balance the mutual distribution of the four subgroups, only 119 from the 518 individuals are selected for the performance test. All remaining images are added to the enrolled images, which causes a different number of enrolled images between MEDS-II and CHD.

6 FACE RECOGNITION AUDIT

In this section the audit is specified. First the datasets, thresholds, and type of assessment are discussed. Then, the configuration of the algorithm is explained. At last, the results of the audit are described.

	MEDS-II	CHD	Total
All Images	1300	628	1928
Enrolled Images	1181	527	1708
Input Images	119	101	220
Males	73	61	134
Females	46	40	86
Total Black	60	41	101
Black Males	37	21	58
Black Females	23	20	43
Total White	59	60	119
White Males	36	40	76
White Females	23	20	43
Total Headgear	0	101	101
Headscarf	0	73	73
Cap	0	10	10
Hat	0	18	18

Figure 9: Overview of specified datasets

6.1 Audit Specifications

Two types of images emerged in face recognition processes: input images and enrolled images. The Rekognition algorithm uses an input image to scan the identity. The similarity score is determined based on the similarities between the input image's identity and enrolled identities. A similarity score is within a range of 0% and 100%.

6.1.1 Datasets. For investigating the effects of headgear on the performance of the Amazon Rekognition algorithm, the audit centralizes the results of the CHD. The CHD includes 527 enrolled images and 101 input images. The audit generates 53.227 similarity scores for processing the CHD. In order to assess the effects of headgear, the results of the CHD are compared with the results of the MEDS-II, as benchmark. The MEDS-II includes 119 input images and 1181 enrolled images. For the MEDS-II, the audit generates 128.639 similarity scores. In total the audit generates 181.866 similarity scores.

6.1.2 Thresholds. A threshold for each dataset is determined based on the distribution of the similarity scores. The threshold is set using a percentile score for mated and non-mated results. The mated results are the scores with matching images of the input image. The threshold is set based on the average value of the 10th percentile for mated results and the 90th percentile for non-mated results. The 90th percentile is the value where 90% of all non-mated similarity scores are fall at or below the percentile's value. The 90th percentile of CHD's non-mated results is 22,97%, which means that 90% of the non-mated results are at or below 22,97%. The 10th percentile of CHD's mated results is 88,87%. The threshold is set on the average of these values, which is 55,92%.

By picking the average of the two percentile values, the thresholds are set per dataset according to the results of the algorithm. Only the results that are not located in the 10th and 90th percentile for both outcomes may exceed the threshold and will be qualified as false positives or false negatives. Setting a threshold according to the distribution of the results, ensures reliable performance of the algorithm. Establishing a threshold randomly would indicate misleading outcomes because the threshold is not adjusted to the results of the dataset. It could generate many false positives and false negatives because the threshold is not aligned with the distribution of the results.

For MEDS-II, the 90th percentile of MEDS-II's non-mated results is 23,93%, and the 10th percentile for mated results is 97,68%. By averaging the 10th and 90th percentile, the threshold of the MEDS-II is 60,81%.

6.1.3 Results. To assess the results of the CHD, the components of the confusion matrix are applied. As described in 3.3, the confusion matrix contains four categories: true positives, true negatives, false positives, and false negatives. Results are categorized in one of the four components by checking if the predicted value is similar to the actual value. With the threshold of 55,92% for the CHD, the mated results of the CHD are qualified as either true positive or false negative, and the non-mated results are qualified as either true negative of false positive. Based on the categorization of all results, as described in 3.3, the PPV, FDR, TPR, and FPR are calculated. The same process is executed for MEDS-II independently in order to compare the results between the two datasets.

6.2 Algorithm Configuration

The Amazon Rekognition algorithm is part of Amazon Web Services (AWS). AWS is an online platform for cloud computing. AWS provides on-demand capabilities for data storage and computing power to offer cloud-based services like Amazon Rekognition and Amazon Transcribe [18]. The functions of the Amazon Rekognition algorithm are accessible via the Rekognition API. Images need to be shared with the Rekognition API to execute the face recognition tasks [19].

To make the images accessible for Amazon Rekognition algorithm, images are stored in Amazon's Simple Storage Service (Amazon S3) [20]. Amazon S3 stores the images in an S3 bucket. Every S3 bucket has a unique identifier. The Rekognition API requires an S3 identifier to access and process all images. The process of face recognition with AWS contains three components: a python program, the Amazon Face Rekognition algorithm, and Amazon S3. The python program is written in python3 [21] and enables face recognition functions. Changes in the AWS keys and configurable elements like the threshold are adapted concerning the specifications of the performance test. In this research, the original command names are changed into names to make them readily understandable. The original command names were numbers (1,2,3, and 4) and are changed into create_collection, index_faces, search_faces,

[18] https://aws.amazon.com/what-is-aws/
[19] https://aws.amazon.com/rekognition/
[20] https://aws.amazon.com/s3/
[21] https://radiostud.io/machine-learning-use-case-facial-recognition-using-amazon-rekognition/

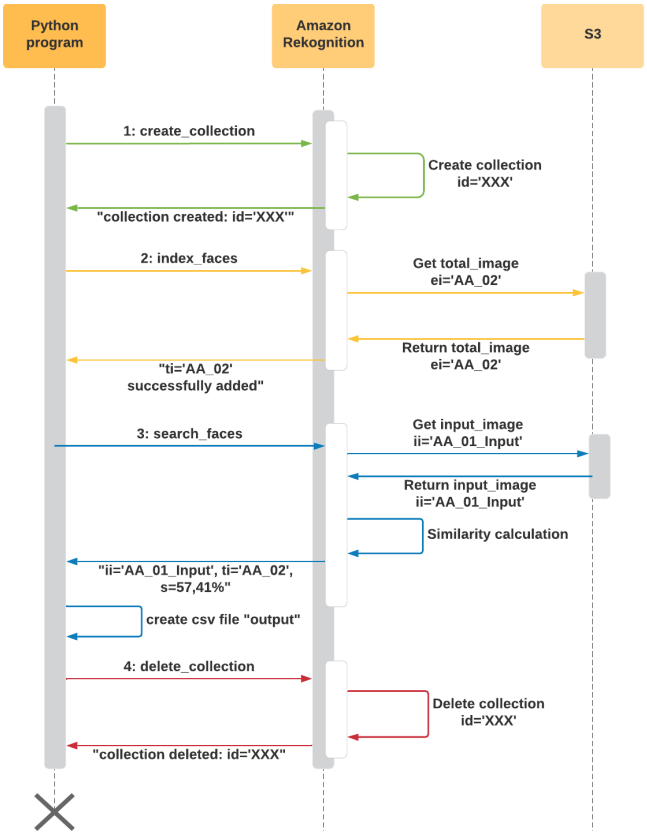


Figure 10: UML Diagram of Face Recognition Process

and delete_collection. The python program used for the audit is attached in A. For running the python program, an AWS account with full access to Amazon's face recognition services is required. When all images are stored in an S3 bucket, the python program needs to be configured. Every AWS account includes unique access and secret keys. The access and secret key need to be adapted in the python program. The "threshold" and "maxFaces" are changeable in the python program according to the requirements of the task. Every function of the python program has a unique command name which is executable via the terminal. Figure 10 presents a UML-diagram of the face recognition process. The UML-diagram presents the sequence of steps per function. The UML-diagram shows the entire process for calculating a similarity score based on the similarities between input image (ii) AA_01_Input and enrolled image (ei) AA_02. The output file of the Amazon Rekognition algorithm includes three columns: the ID of the input, the ID of the enrolled image, and the similarity score.

In total, the python program contains four commands. The four commands are:

- **create_collection**: The command create_collection creates a virtual collection. This virtual collection is required for storing facial data of the images in the S3 bucket. The create collection process is presented in figure 10 with green-colored lines.

Dataset	Metrics	Overall	Headscarf	Cap	Hat
CHD	PPV	99,55%	99,45%	99,77%	99,83%
	FDR	0,45%	0,55%	0,23%	0,17%
	TPR	97,61%	97,15%	100%	98,15%
	FPR	0,43%	0,53%	0,23%	0,16%
	TP	3,14	3,08	3,10	3,39
	TN	521,49	521,21	522,70	522,72
	FP	2,28	2,78	1,20	0,83
	FN	0,10	0,12	0,00	0,06

Figure 11: Performance Results CHD regarding Headgear

Dataset	Metrics	Overall	Headscarf	Cap	Hat
MEDS-II	PPV	99,82%	-	-	-
	FDR	0,18%	-	-	-
	TPR	98,12%	-	-	-
	FPR	0,18%	-	-	-
	TP	1,60	-	-	-
	TN	1177,26	-	-	-
	FP	2,13	-	-	-
	FN	0,02	-	-	-

Figure 12: Performance Results MEDS-II regarding Headgear

- **index_faces**: The command index_faces generates facial data of every image in the S3. This data gets stored in the virtual collection. When a process is successfully executed, the python program returns 'image_id: success'. After processing all images, the system is ready to execute the face recognition tasks. The index faces process is presented in figure 10 with yellow-colored lines.
- **search_faces**: The command search_faces scans the image IDs that are located in the local directory input_images, to list the input images. A similarity score is calculated by the Face Rekognition algorithm based on the similarities between the input image and every image in the virtual collection. After processing, the python program provides a CSV file with the results that meet the requirements that are configured in the python program. The search faces process is presented in figure 10 with blue-colored lines.
- **delete_collection**: The command delete_collection removes the existing collection including all indexed data. The delete collection process is presented in figure 10 with red-colored lines.

6.3 Audit Results

In this section, the audit results are summarized. First, the results of the CHD concerning headgear are defined. The audits results of the CHD regarding headgear are visible in figure 11. Then, in comparison with the results of the MEDS-II in figure 12, the differences between the CHD and MEDS-II are described. Figure 13 compares the overall results of the two datasets. In the Appendix, a specification on race and gender of the audit results per dataset is visible. A specification of the CHD and MEDS-II is apparent in B and C respectively.

Metrics	Headgear	No Headgear	Difference
	CHD	MEDS-II	
PPV	99,55%	99,82%	-0,27%
FDR	0,45%	0,18%	0,27%
TPR	97,61%	98,12%	-0,51%
FPR	0,43%	0,18%	0,25%

Figure 13: Comparison Performance Results

6.3.1 Results. The PPV of the CHD is set on 99,55% with an FDR of 0,45%. The Amazon Face Rekognition algorithm returns 2,28 false positives and 0,10 false negatives per search on average. Looking at the results per gender of the CHD, the FDR is 0,33% and 0,64% for males and females respectively. The PPV shows a difference of 0,31% and the TPR a difference of 1,55%, both negatively for females. The Amazon Rekognition algorithms return per search on average 3,2 false positives for females and 1,67 false positives for males.

From the three types of headgear, the PPV of headscarves is the lowest. The PPV of hats holds the highest PPV with 99,83%. The FPR of headscarves is the biggest among the three types with 0,53%. On average, an image with a headscarf returns 2,78 false positives per search. The average number of false positives is 1,20 for images with caps.

Among the four subgroups of race and gender, the black males are the most accurate with a score of 99,76%. The lowest PPV is measured on the black females, with a score of 99,03%. The FDR on black females is set on 0,97%. The other three subgroups have an FDR between 0,24% and 0,37%. On average, a black female search returns 5,05 false positives. The number of false positives on white males, white females, and black males is 1,88, 1,35, and 1,29, respectively. The number of false negatives is the highest on white females with 0,25 false positives per search.

6.3.2 CHD versus MEDS-II. Figure 13 presents the differences in the overall performance results between CHD and MEDS-II. The PPV of the MEDS-II is set on 99,82%, which is 0,27% higher than the CHD. The TPR is measured on 98,12% for MEDS-II, 0,51% higher than the TRP of CHD.

B specifies the overall performance results of the CHD. The specification of the results of MEDS-II is presented in C. Among the subgroups, the error metrics are higher for CHD. The Amazon Rekogntion algorithm performs less on white males, white females, and black females of the CHD in comparison with MEDS-II. In contrast to the CHD, the results of the MEDS-II are in favor of females. The PPV on males is 99,76%, which is 0,16% lower than females. For the MEDS-II, the FDR of males is three times smaller than females. The Amazon Rekognition algorithms return per search on average 0,93 false positives for females and 2,88 false positives for males. The FDR on females is for the CHD eight times higher than the FDR on females of MEDS-II. The FDR on males of the CHD is 0,09% higher than the FDR on males of the MEDS-II.

The Amazon Rekognition algorithm indicates, in comparison with the MEDS-II results, lower performance on the CHD. The FDR is on average 2,5 times higher for the CHD. Despite the doubled number of enrolled images and a higher threshold for the MEDS-II, the number of false positives of black females is four times more on the CHD. The FPR for a black female search is on average nine times higher for the CHD.

7 AUDIT ANALYSIS

The definition of fairness in this research, as defined in 3.2, contains conditional requirements. The conditional requirements of fairness are a false positive error rate of 0,1%, and a positive predictive value of 99,80%. These conditional requirements of fairness are based on the performance of forensic fingerprints. Furthermore, the definition of fairness requires predictive parity and predictive equality, which aim for an equal value of the positive predictive value (PPV) and false positive rate (FPR) among the subgroups of the CHD. In the extension of predictive parity and predictive equality, the results of the CHD are compared with the scores of MEDS-II, as a benchmark. Based on the fulfillment of these criteria of fairness, the algorithmic fairness of the Amazon Rekognition algorithm is ascertained. The ethical implications are examined according to the usage in law enforcement, based on whether the Amazon Rekognition algorithm is processing headgear fairly.

Conditional Requirements. The overall performance results of the CHD does not fulfill the conditional requirements of the PPV and FPR. According to the conditional requirements of fairness, the overall PPV of the CHD is 0,25% too low, and the overall FPR is 0,33% too high. The specified results on race and gender of the CHD, noticeable in B, does not achieve the conditional requirements of the PPV and FPR. Notably, the PPV and FPR of black females with headgear have the most significant divergence from the conditional requirements.

Predictive Parity. The predictive parity requires a similar PPV among the groups within the same dataset. Referring to the specified results in B, the PPV varies in the CHD among the types of headgear, gender and race. None of the PPVs has the same score. Among the white males, white females, black males, and black females, the PPV of the CHD fluctuate in a range of 0,67%, from 99,03% to 99,76%. The PPV of the CHD also differs from the PPV of the MEDS-II. The overall PPV is 0,27% lower for the CHD. Among the subgroups, the PPV of the CHD is lower than the MEDS-II on three of the four subgroups, especially on black females. The PPV on black females is 0,85% lower of the CHD. The CHD is only performing better on black males in comparison with MEDS-II.

Predictive equality. Additional to predictive parity is the investigation of predictive equality. The predictive equality requires a similar FPR among the groups within the same dataset. The FPR differs among the subgroups. The FPR of the CHD ranges from 0,24% for black males until 0,97% for black females. The differences in the FPR induce different numbers of false positives per subgroup. A black female search returns on average 5,05 false positives, which is four times more than the number of false positives for a black male search. The FPR of the CHD also diverges from the FPR of the MEDS-II. The FPR is 0,25% higher for the CHD. The difference is expressly visible in the average number of false positives per black female search. A black female search of the CHD returns four times more false positives than a black female search of MEDS-II.

7.1 Algorithmic Fairness

The performance of the Amazon Rekognition algorithm does not achieve the conditional requirements of fairness in processing the CHD. The FPR of the CHD is 0,33% too high, and the PPV of the CHD is 0,25% too low. Furthermore, the performance results of the CHD do not fulfill the requirements of predictive parity and predictive equality. The PPV and FPR differ among the subgroups. Moreover, in comparison with MEDS-II as a benchmark, the overall results differ between the datasets. The results of the CHD specified on gender or race, are worse than the MEDS-II results. The deviation between the CHD and MEDS-II means that the Amazon Rekognition algorithm is performing worse on people wearing headgear. As the results of the CHD by the Amazon Rekognition algorithm does not achieve the conditional requirements of fairness, the predictive parity, and predictive equality, the algorithm is claimed as unfair.

Regarding the function of deep neural network (DNN) as described in 2.1, and the impact of defects in the training data as described in 3.1, the flaws of the Amazon Rekognition algorithm in processing the CHD could be caused by a lack of images with headgear in the data on which the algorithm is trained or through implementation of biased behavior during development. The application of erroneous labels in training data could also induce flaws concerning headgear. Due to this shortage in the data on which the Amazon Rekognition is trained, incorrect weights are produced. The weights are adjusted and corrected on faulty labels. The lack of performance by the Amazon Rekognition algorithm on people with headgear indicates headgear bias.

7.2 Ethical Implications

The likelihood of generating false positives in face recognition tasks on people with headgear is not achieving the conditional requirement of fairness. A false positive could classify an innocent person as a suspected criminal. The discovery of unfair behavior in the Amazon Rekognition algorithm induces ethical implications. The unfair behavior implies a limited performance for face recognition tasks in processing people with headgear. The performance assessment indicates headgear bias as the algorithms show an overall preference for people wearing no headgear. Disadvantaging people with headgear implies discriminative outcomes. The contemporary usage of the same algorithm in law enforcement by police departments in the U.S. will produce misidentifications and discriminative outcomes. The algorithm processes surveillance footage for the identification of suspected people [11, 25]. A misidentification in law enforcement leads to the arrest of an innocent person, while the actual perpetrator remains free. This scenario is ethically not permitted because it violates the individual's freedom and rights.

8 DISCUSSION

The performance results of the Amazon Rekognition algorithm on the CHD indicates headgear bias. Erroneous identifications are likely to occur in law enforcement on people who are wearing headgear. Erroneous identifications could lead to the arrest of innocent people, while the real perpetrator remains free. The arrest of an innocent person is not ethically allowed as it violates the individual's freedom and rights. Since police departments are currently adopting the same face recognition technology to support law enforcement, the occurrence of headgear bias should urge police departments to reconsider their usage of the Amazon Rekognition algorithm. At this state, human intervention is required to prevent misidentifications in law enforcement. Further improvement of the Amazon Rekogntion algorithm in processing people with headgear is required.

Extension of the CHD by involving more images could substant the statements of this research. More images would generate more results and could emphasize the challenges of the CHD. Additionally, as the CHD is currently mainly focused on headscarves, expansion of the CHD could focus on involving various headgear from other religions.

The performance test of the Amazon Rekognition algorithm intends to imitate the real usage of the algorithm in law enforcement in the U.S. However, detailed information regarding the type of process, conditional requirements, and technical configurations are missing. Therefore, the occurrence of unfair behavior of the Amazon Rekognition algorithm is determined generally, not exclusively stated upon the single-use case within law enforcement.

Face recognition tasks include several challenges. Differences between the input image and the enrolled image caused by illumination, pose, aging, facial expression, or occlusion, makes face recognition hard. The CHD includes images from various circumstances like movie scenes, catwalks, sports matches, and model shots. Additionally, CHD contains images from the same identity with and without headgear. People who are wearing a headscarf for their religion are wearing a headscarf constantly. As face recognition on people with and without headscarf is increasing the complexity of face recognition, an extension of the CHD by adding images of individuals only with headgear could also focus on increasing the imitation of real-life usage.

Currently, the meta documentation of the CHD is manually annotated exclusively by the creator himself. Verification of the annotations by external reviewers could improve the quality of the annotations. Extension of the meta documentation could make CHD suitable for other purposes like age determination and facial coordinates detection.

Through this research, the Amazon Rekognition algorithm is the first algorithm that has processed the CHD. Other face recognition algorithms must be tested on datasets that include headgear. Multiple performance results enable mutual comparison of the results, and could specify the challenges of the CHD.

9 CONCLUSION

This research obeys the necessity for testing the performance of face recognition algorithms on the appearance of headgear. The CHD is the first dataset that introduces the opportunity to test the performance of face recognition exclusively on headgear. The performance test with the CHD ascertains unfair behavior of the Amazon Rekognition algorithm. The Amazon Rekognition algorithm flaws in processing people with headgear based on the conditional requirements of fairness. Therefore, a headgear bias is considered within the Amazon Rekognition algorithm. Headgear bias implies a higher likelihood for misidentifications in law enforcement that could lead to the arrest of innocent people, while the real perpetrator remains free. The arrest of an innocent person is not ethically

permitted as it violates the individual's freedom and rights. The higher probability of misidentifications for people with headgear in comparison with people wearing no headgear induces unequal treatment of people. Unequal treatment indicates discriminative behavior of the Amazon Rekognition algorithm, which is not permitted according to the first article of the Dutch constitution. As face recognition enhances law enforcement, it is not aimed to doubt the development of face recognition, but to indicate the current performance. The performance test in this research seeks to prior the limitations and effects of face recognition algorithms. Further research should investigate, test, and attempt to solve headgear bias. At the same time, police departments should reconsider their actual usage of the Amazon Rekognition algorithm in law enforcement.

REFERENCES

[1] Kasun Amarasinghe, Kevin Kenney, and Milos Manic. 2018. Toward explainable deep neural network based anomaly detection. In *2018 11th International Conference on Human System Interaction (HSI)*. IEEE, 311–317.

[2] Alan Joseph Bekker and Jacob Goldberger. 2016. Training deep neural-networks based on unreliable labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2682–2686.

[3] Joanna Bryson and Alan Winfield. 2017. Standardizing ethical design for artificial intelligence and autonomous systems. *Computer* 50, 5 (2017), 116–119.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[5] Beverly Chico. 2000. Gender headwear traditions in Judaism and Islam. *Dress* 27, 1 (2000), 18–36.

[6] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[7] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *interactions* 25, 6 (2018), 58–63.

[8] C Dolman and D Semenovich. [n. d.]. Algorithmic Fairness: Contemporary Ideas in the Insurance Context. ([n. d.]).

[9] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

[10] Andrew P Founds, Nick Orlans, Whiddon Genevieve, and Craig I Watson. 2011. *Nist special databse 32-multiple encounter dataset ii (meds-ii)*. Technical Report.

[11] Clare Garvie. 2016. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology.

[12] Adrian Iustin Georgevici and Marius Terblanche. 2019. Neural networks and deep learning: a brief introduction.

[13] Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. 2018. *Ongoing face recognition vendor test (frvt) part 2: Identification*. Technical Report.

[14] M Hassaballah and Saleh Aly. 2015. Face recognition: challenges, achievements and future directions. *IET Computer Vision* 9, 4 (2015), 614–626.

[15] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801.

[16] Brais Martinez and Michel F Valstar. 2016. Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in face detection and facial image analysis*. Springer, 63–100.

[17] Rajakishore Nath and Vineet Sahu. 2017. The problem of machine ethics in artificial intelligence. *AI & SOCIETY* (2017), 1–9.

[18] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[19] Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemendy, and Z Bauer. 2018. The AI Index 2018 Annual Report. *AI Index Steering Committee, Human-Centered AI Initiative, Stanford University. Available at http://cdn. aiindex. org/2018/AI% 20Index* 202018 (2018).

[20] Songül Tolan. 2019. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges. *arXiv preprint arXiv:1901.04730* (2019).

[21] Jim Torresen. 2018. A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI* 4 (2018), 75.

[22] Bradford T Ulery, R Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. 2011. Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences* 108, 19 (2011), 7733–7738.

[23] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[24] Sofia Visa, Brian Ramsay, Anca L Ralescu, and Esther Van Der Knaap. 2011. Confusion Matrix-based Feature Selection. *MAICS* 710 (2011), 120–127.

[25] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI now report 2018*. AI Now Institute at New York University.

[26] Ming-Hsuan Yang, David J Kriegman, and Narendra Ahuja. 2002. Detecting faces in images: A survey. *IEEE Transactions on pattern analysis and machine intelligence* 24, 1 (2002), 34–58.

[27] Serena Yeung, N Lance Downing, Li Fei-Fei, and Arnold Milstein. 2018. Bedside Computer Vision-Moving Artificial Intelligence from Driver Assistance to Patient Safety. *The New England journal of medicine* 378, 14 (2018), 1271.

[28] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. 2015. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding* 138 (2015), 1–24.

10 APPENDIX

A PYTHON PROGRAM

```
# -*- coding: utf-8 -*-
"""
Created on Tue Oct  9 14:24:00 2018

@author: SHREYA
"""

import boto3
from botocore.exceptions import ClientError
import sys
import re
import os

collectionId='image-collection2'#Creating a collection
bucket='chd2906__init__'

threshold = 0
maxFaces = 1400

client=boto3.client('rekognition', aws_access_key_id = 'X
    ', aws_secret_access_key = 'X', region_name = 'X')



def index_faces(photo):       #adding faces to the
    collection
    try:
        response = client.index_faces(CollectionId=
    collectionId,
                                      Image={'S3Object':{'
    Bucket':bucket,'Name':photo}},
                                      ExternalImageId=photo,
                                      MaxFaces=1,
                                      QualityFilter="AUTO",
                                      DetectionAttributes=['ALL
    '])
        return response['FaceRecords']

    except ClientError as e:

        print("Boto3 Error : " + e.response['Error']['
    Code'])
        return None

def search_faces_by_image(photo):     #searching faces
    from the collection

    try:
```

```
        response = client.search_faces_by_image(CollectionId=collectionId,
                                    Image={'S3Object':{'Bucket':bucket,'Name':photo}},
                                    FaceMatchThreshold=threshold,
                                    MaxFaces=maxFaces)
        return response['FaceMatches']

    except ClientError as e:

        print("Boto3 Error : " + e.response['Error']['Code'])
        return None

def create_collection():
    try:
        response=client.create_collection(CollectionId=collectionId)
        print("collection created")
        return response['StatusCode']
    except ClientError as e:

        print("Boto3 Error : " + e.response['Error']['Code'])
        return None

def delete_collection():
    try:
        response =client.delete_collection(CollectionId=collectionId)
        print("collection deleted")
        return response['StatusCode']

    except ClientError as e:

        print("Boto3 Error : " + e.response['Error']['Code'])
        return None

def check_result(par,response,photo=None):

    if response == None:
        print("operation failed")

    else:
        if par == '1':

            if(response == 200):
                print('Successfully created the collection ' + collectionId)
            else:
                print("Collection creation failed")


        elif par == '2':

            if(response == 200):
                print('Successfully deleted the collection ' + collectionId)
            else:
                print("Collection deletion failed")

        elif par == '3':

            for faceRecord in response:
                print('Results for ' + photo + ':')
                print('Face ID : ' + faceRecord['Face']['FaceId'])
                print('Location : {}'.format(faceRecord['Face']['BoundingBox']))
                print('Photo is uploaded to the collection :' + collectionId)

        elif par == '4':

            for match in response:
                print('FaceId :' + match['Face']['FaceId'])
                print('Similarity :' + "{:.2f}".format(match['Similarity']) + "%")
                print('The given ' + photo + ' matches with: '+ match['Face']['ExternalImageId'])


if __name__ == '__main__':

    if len(sys.argv) == 2:
        operation = sys.argv[1]
        counter = 0

        if operation == "create_collection":
            create_collection()

        if operation == "delete_collection":
            delete_collection()

        if operation == "index_faces":
            total_images = os.listdir("./total_images")
            for ti in total_images:
                counter = counter +1
                if ti.endswith(".jpg") or ti.endswith(".png") or ti.endswith(".jpeg"):
                    response = index_faces(ti)
                    if(None == response):
                        print("Indexing failed")
                    else:
                        print("Success "+ str(counter) +" : " + ti)

        if operation == "search_faces":
            # Defineer .csv headers
            text_file = open("output.csv", "a")
            text_file.write("ii,ti,s\n")
            # Defineer input_images
            input_images = os.listdir("./input_images")

            # Vergelijk images met Collection
            for ii in input_images:
                counter = counter + 1
                if ii.endswith(".jpg") or ii.endswith(".png") or ii.endswith(".jpeg"):
                    response_two = search_faces_by_image(ii)
                    if(None == response_two):
                        print("Searching failed")
                    else:
                        print("Success "+ str(counter) +" : " + ii)
                        for match in response_two:
                            # Schrijf naar csv file
                            s = "{:.2f}".format(match['Similarity']) + "%"
                            text_file.write(ii+","+match['Face']['ExternalImageId']+","+s+"\n")

            text_file.close()
```

```
157
158     # if len(sys.argv) > 1:
159
160     #     operation_select =sys.argv[1]
161     #     response_obj = None
162
163     #     if len(sys.argv) == 2:
164     #         if operation_select =='1':
165     #             response_obj = create_collection()
166
167     #         elif operation_select == '2':
168     #             response_obj = delete_collection()
169
170     #         check_result(operation_select, response_obj
        )
171
172     #     elif len(sys.argv) == 3:
173
174     #         pic_path = sys.argv[2]
175
176     #         #if re.match('[a-z]+\-[0-9]+\.jpg',pic_path
        ):
177     #         if pic_path.endswith('.jpg'):
178
179     #             if operation_select == '3':
180     #                 response_obj = index_faces(pic_path
        )
181     #                 if(None == response_obj):
182     #                     print("Indexing failed")
183
184
185
186     #             elif operation_select == '4':
187     #                 response_obj =
         search_faces_by_image(pic_path)
188     #                 if(None == response_obj):
189     #                     print("Searching failed")
190
191     #             check_result(operation_select,
         response_obj,pic_path)
192
193     #         else:
194     #             print("Incorrect image file format,
        expected .jpg")
195
196
197
198     #     else:
199     #         print("Insufficient arguments")
200
201
202
203     # else:
204     #     print("Insufficient arguments")
205
206
207
208
209
210
211
212
213     #
```

B RESULTS CHD ON GENDER AND RACE

Dataset	Metrics	F	M	B	W	WM	WF	BM	BF
CHD (With Headgear)	PPV	99,36%	99,67%	99,40%	99,65%	99,63%	99,70%	99,76%	99,03%
	FDR	0,64%	0,33%	0,60%	0,35%	0,37%	0,30%	0,24%	0,97%
	TPR	96,67%	98,22%	99,19%	96,53%	97,29%	95,00%	100%	98,33%
	FPR	0,61%	0,32%	0,60%	0,32%	0,36%	0,26%	0,25%	0,96%
	TP	3,05	3,20	3,27	3,05	3,03	3,10	3,52	3,00
	TN	520,60	522,07	520,59	522,10	522,00	522,30	522,19	518,90
	FP	3,20	1,67	3,12	1,70	1,88	1,35	1,29	5,05
	FN	0,15	0,07	0,02	0,15	0,10	0,25	0,00	0,05

C RESULTS MEDS-II ON GENDER AND RACE

Dataset	Metrics	F	M	B	W	WM	WF	BM	BF
MEDS-II (No Headgear)	PPV	99,92%	99,76%	99,76%	99,88%	99,82%	99,96%	99,69%	99,88%
	FDR	0,08%	0,24%	0,24%	0,12%	0,18%	0,04%	0,31%	0,12%
	TPR	94,20%	100,00%	97,22%	99,05%	100,00%	97,22%	100,00%	90,91%
	FPR	0,08%	0,24%	0,24%	0,12%	0,18%	0,04%	0,31%	0,12%
	TP	1,22	1,84	1,58	1,61	1,89	1,17	1,78	1,26
	TN	1178,80	1176,29	1176,62	1177,92	1177,03	1179,30	1175,57	1178,30
	FP	0,93	2,88	2,78	1,46	2,08	0,48	3,65	1,39
	FN	0,04	0,00	0,02	0,02	0,00	0,04	0,00	0,04