PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

CAMILA KOLLING DOS REIS

MITIGATING BIAS IN FACIAL ANALYSIS SYSTEMS BY INCORPORATING LABEL DIVERSITY

Porto Alegre
2022

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

MITIGATING BIAS IN FACIAL ANALYSIS SYSTEMS BY INCORPORATING LABEL DIVERSITY

CAMILA KOLLING DOS REIS

Master Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. Dr. Soraia Raupp Musse
Co-Advisor: Prof. Dr. Adriano Alonso Veloso

Porto Alegre
2022

Ficha Catalográfica

K81m Kolling, Camila

Mitigating Bias in Facial Analysis Systems by Incorporating Label Diversity / Camila Kolling. – 2022.
97 p.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Soraia Raupp Musse.
Coorientador: Prof. Dr. Adriano Alonso Veloso.

1. Deep Learning. 2. Facial Analysis. 3. Neural Networks. 4. Fairness. I. Musse, Soraia Raupp. II. Veloso, Adriano Alonso. III. , . IV. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Loiva Duarte Novak CRB-10/2079

CAMILA KOLLING DOS REIS

MITIGATING BIAS IN FACIAL ANALYSIS SYSTEMS BY INCORPORATING LABEL DIVERSITY

This Master Thesis has been submitted in partial fulfillment of the requirements for the degree of Master in Computer Science, of the Computer Science Graduate Program, School of Technology of the Pontifical Catholic University of Rio Grande do Sul

Sanctioned on August 04, 2022.

COMMITTEE MEMBERS:

Prof. Dr. Virgílio Augusto Fernandes Almeida (PPGCC/UFMG)

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS)

Prof. Dr. Adriano Alonso Veloso (PPGCC/UFMG- Co-Advisor)

Prof. Dr. Soraia Raupp Musse (PPGCC/PUCRS - Advisor)

I dedicate this work to my parents.

“Remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the universe exist. Be curious. And however difficult life may seem, there is always something you can do and succeed at. It matters that you don’t just give up.”
(Stephen Hawking)

ACKNOWLEDGMENTS

I, Camila Kolling dos Reis, would like to especially thank my parents, who have always supported and encouraged me during all the stages of my life. My family provided me with the opportunity to study and enhance my personal and professional abilities, and I am eternally grateful for that.

I would also like to thank all the colleagues and friends who were present during my Master's, all of whom were of great help and importance. Special thanks to Victor Araujo, who joined all the past years' meetings and spent long afternoons debating research with me.

I would like to thank the company Kunumi, which provided me the opportunity of making my life through research. I would also like to thank CAPES, who financially supported my Master's.

Thank you to the committee members Prof. Dr. Rafael Bordini and Prof. Dr. Virgílio Fonseca, for their availability and patience in evaluating my work.

Special thanks to my advisors, Prof. Dr. Soraia Raupp Musse and Prof. Dr. Adriano Afonso Veloso, who trusted, helped, and guided me throughout the development of this work. Nothing of this would have been possible without your support!

MITIGANDO VIÉS EM SISTEMAS DE ANÁLISE FACIAL AO INCORPORAR DIVERSIDADE DE RÓTULO.

RESUMO

Modelos de análise facial são cada vez mais utilizados em aplicações do mundo real que têm impacto significativo na vida das pessoas. No entanto, como demonstrado pela literatura, os modelos que classificam automaticamente os atributos faciais podem apresentar comportamento de discriminação em relação a grupos protegidos, potencialmente causando impactos negativos nos indivíduos e na sociedade. Portanto, é fundamental desenvolver técnicas que possam mitigar vieses não intencionais em classificadores faciais. Assim, neste trabalho, apresentamos um novo método de aprendizado de máquina que combina rótulos subjetivos, baseados em humanos, e anotações objetivas, baseadas em definições matemáticas, de traços faciais. Especificamente, geramos novas anotações objetivas a partir de dois conjuntos de dados anotados por humanos em grande escala, cada um capturando uma perspectiva diferente do traço facial analisado. Em seguida, propomos um método de aprendizado em conjunto, que combina modelos individuais treinados em diferentes tipos de anotações. Fornecemos uma análise aprofundada do procedimento de anotação, bem como a distribuição dos conjuntos de dados. Além disso, demonstramos empiricamente que, ao incorporar a diversidade de rótulos, nosso método mitiga com sucesso vieses não intencionais, mantendo uma precisão significativa nas tarefas.

Palavras-Chave: aprendizado profundo, análise facial, redes neurais, justiça.

MITIGATING BIAS IN FACIAL ANALYSIS SYSTEMS BY INCORPORATING LABEL DIVERSITY

ABSTRACT

Facial analysis models are increasingly applied in real-world applications that have significant impact on peoples' lives. However, as previously shown, models that automatically classify facial attributes might exhibit algorithmic discrimination behavior with respect to protected groups, potentially posing negative impacts on individuals and society. It is therefore critical to develop techniques that can mitigate unintended biases in facial classifiers. Hence, in this work, we introduce a novel learning method that combines both subjective human-based labels and objective annotations based on mathematical definitions of facial traits. Specifically, our proposed method first generates new objective annotations, each capturing a different mathematical perspective of the analyzed facial traits. We then use an ensemble learning method, which combines individual models trained on different types of annotations. We provide an in-depth analysis of the annotation procedure as well as the datasets distribution. Moreover, we empirically demonstrate that, by incorporating label diversity to the decision-making process, our method successfully mitigates unintended biases, while maintaining significant accuracy on the downstream tasks.

Keywords: deep learning, facial analysis, neural networks, fairness.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI. – Artificial Intelligence

AUs. – Action Units

CelebA. – CelebFaces Attributes Dataset

CFD. – Chicago Face Database

DL. – Deep Learning

ExpW. – Expression in-the-Wild Database

FACS. – Facial Action Coding System

FER. – Facial Expression Recognition

FNR. – False Negative Rate

FPR. – False Positive Rate

GANs. – Generative Adversarial Networks

GDPR. – General Data Protection Regulation

GR. – Golden Ratio

ITA. – Individual Topology Angle

kNN. – k-Nearest Neighbors

LCS. – Longest Common Subsequence

ML. – Machine Learning

NC. – Neoclassical Canons

Sym. – Symmetry

TNR. – True Negative Rate

TPR. – True Positive Rate

LIST OF SYMBOLS

CONTENTS

1. INTRODUCTION

In recent years, artificial intelligence (AI) has been incorporated into a large number of real-world applications, such as multimedia concept retrieval (Pouyanfar et al., 2019), image classification (Loussaief and Abdelkrim, 2016), video and product recommendation (Das et al., 2017), and social network analysis (Ghani et al., 2019). These data-driven systems increasingly surround and shape humans in their daily life. This is the case especially for machine learning (ML) algorithms, which learn human behavior by recognizing patterns within the existing data and applying them to new unseen instances to make predictions for future outcomes (Goodfellow et al., 2016). There are several advantages that foster the adoption of such systems (Kleinberg et al., 2018). Firstly, algorithms may process much more data, and thus take into account much more factors and context, than human beings may hold. Secondly, algorithms perform mathematically complex computations much faster than humans, and might even reduce the factor of human error. Finally, human decisions are subjective, and they often include biases.

Therefore, by integrating automated algorithm-based decision-making systems one might expect that the decisions will be more objective and fair, especially for high-stake applications, including employment, criminal justice, and personalized medicine. Unfortunately, as several approaches have shown (Buolamwini and Gebru, 2018; Bolukbasi et al., 2016; Caliskan et al., 2017), this is not always the case. Given that AI algorithms are trained with historical data, these prediction engines may inherently learn, preserve and even amplify these biases (Kleinberg et al., 2016; Zhao et al., 2017). As defined by Mehrabi et al. (2021), in the context of algorithmic decision-making, “fairness is the *absence* of any prejudice or favoritism towards an individual or group based on their inherent or acquired characteristics”. In other words, an algorithm that has skewed decision towards a particular group of people is considered unfair (Mehrabi et al., 2021).

Special attention has been devoted to facial analysis applications since in the biometric modality performance differentials mostly fall across points of sensitivity (*e.g.*, race and gender) (Drozdowski et al., 2021). Particularly, the human face is a very important research topic as it transmits plenty of information to other humans, and thus possibly to computer systems (Mehrabian, 2017), such as identity, intentions, emotional state, attractiveness, age, gender, attention and personality traits. Additionally, the human face has also been of considerable interest to researchers due to the inherent and extraordinarily well-developed ability of humans to process, recognize, identify, and extract information from others’ faces (Little, 2014). Faces dominate our daily situations since we are born, and our sensitivity to faces is strengthened every time we see a face under different conditions (Little, 2014). For instance, previous work show that the responsiveness of human infants only minutes old is greater to face-like stimuli compared with equally complicated non-face stimuli (Goren et al., 1975; Johnson et al., 1991).

With the vast adoption of automated systems in high-impact domains, it is important to take fairness issues into consideration and ensure sensitive attributes will not be used for discriminatory purposes. In this work, we focus our attention on two aspects of facial analysis, *attractiveness* classification and *facial expression* recognition, both of which have a powerful impact on our lives. For instance, the pursuit for beauty has encouraged trillion-dollar cosmetics, aesthetics, and fitness industries, each one promising a more attractive, youthful, and physically fitter version for each individual (Cutler, 2021). Beauty also seems to be an important aspect of human social interactions and matching behaviors, in which more attractive people appear to benefit from higher long-term socioeconomic status and are even perceived by peers as "better" people (Cutler, 2021).

Simultaneously, facial expression is one of the most powerful, instinctive and universal signals for human beings to convey their emotional state and intentions (Darwin, 2014). Often the face express an emotion before people can even understand or verbalize it. Mehrabian and Russell (1974) show that 55% of messages regarding feelings and attitudes are conveyed through facial expression. Facial expression analysis is a fundamental scientific topic, and its study dates back to the Aristotelian era (Bettadapura, 2012).

In part because of its importance and potential uses as well as its inherent challenges, automated attractiveness classification and facial expression recognition have been of keen interest in the computer vision and machine learning communities. Several approaches have proposed methods for automatically assessing face attractiveness and expressions through computer analysis (Ma et al., 2021; Fasel, 2002). However, as already shown in previous work (Sattigeri et al., 2019; Ramaswamy et al., 2021; Chen and Joo, 2021), models that are trained to automatically analyze facial traits might exhibit algorithmic discrimination behavior with respect to protected groups (*e.g.*, gender, race, age), potentially posing negative impacts on individuals and society.

1.1 Goals

Despite several advances towards understanding and mitigating the effect of bias in facial analysis prediction, as far as we are aware, none of the previous work focused on debiasing the system by augmenting the label diversity of the annotations. Thus, in this work, *our goal is to propose and validate an approach that combines different types of annotations, such as the original and subjective human-based labels and the ones we objectively generate based on mathematical definitions of both attractiveness and facial expression*. We hypothesize that introducing *diversity* into the decision-making model by adding mathematical and possibly unbiased notions in the label dimension should reduce the biases present in the final model.

The specific goals of this work involve: (1) review previous literature on debiasing approaches, especially the ones for facial analysis; (2) inspired by previous work, propose a novel method that mitigates fairness issues, while maintaining competitive accuracy; (3) run experiments to test our hypothesis on the proposed method; and (4) analyze results and compare with state-of-the-art approaches. To the best of our knowledge, this is the first time a pre-processing debiasing method combines *objective* (mathematical) labels and *subjective* (human-based) annotations.

1.2 Research Questions

In this work we propose to answer the following questions:

(1) By generating mathematically-based labels (Schmid et al., 2008; Ekman, 1993) and training ML models on these objective annotations, is it possible to obtain a model whose behavior is less discriminatory (measured by a fairness evaluation metric) than one trained on human-based, *i.e.*, subjective, labels?

(2) Is it possible to reduce unfair behavior of facial analysis systems, while maintaining a competitive accuracy, by combining models trained on both subjective and objective notions of attractiveness/facial expressions?

(3) Do the models trained on human-based and mathematically-based labels use the same information to make its final prediction, *i.e.*, do these models attend to the same regions of an image to make its final decision?

1.3 Structure of this Work

The work is organized as follows: Chapter 2 introduces previous works in learning fair models as well as computationally analyzing facial attractiveness and expressions. We also review some works that propose explainability methods for deep learning. We conclude the section by situation our work on the existent liteturature. Chapter 3 describes preliminary notation, datasets, and fairness metrics used throughout this work. We also introduce in detail our proposed method, which is composed of three main parts: (1) generating objective labels for all training datasets; (2) individually training models on both subjective as well as objective annotations; (3) combining individual models into an ensemble. Chapter 4 presents our findings as well as the answers to our research questions. Chapter 5 concludes this work by presenting both ethical as well as our final considerations. We also present future work.

2. RELATED WORK

This chapter aims to cover the main related work concerning fairness, especially in the context of deep learning models, for the tasks explored in this work: facial attractiveness classification and expression recognition. In the following section, we review some of the primary work in fairness regarding deep learning in general (Section 2.1), and also approaches that use ensembles as their main motivation for mitigating fairness issues (Section 2.1.1). We end the first part of this chapter by discussing some of the major evaluation metrics used to mathematically measure and represent bias, fairness, and/or discrimination of machine learning models (Section 2.1.2).

Then, we focus on describing approaches that automatically assess the attractiveness of the face (Section 2.2), including golden ratio (Section 2.2.1), neoclassical canons (Section 2.2.2) and facial symmetry (Section 2.2.3), also covering methods that studied and reduced fairness issues in this task. We next explore previous work on facial expression recognition (Section 2.3), especially in the context of reducing the discriminatory behavior of such systems. Furthermore, we describe the explainability methods used in this work (Section 2.4) to better understand the models that compose our ensemble. We conclude this chapter by situating our work in the literature (Section 2.5).

2.1 Fairness in Deep learning

While fairness is a fairly new topic in machine learning research, it has been extensively studied in the legal and sociological fields (Crenshaw, 1989; Balkin and Siegel, 2003; Small and Pager, 2020). The term discrimination can have two main definitions: (i) *disparate treatment* (Jagielski et al., 2019): treating individuals differently by making use of the protected attribute in the decision-making process (*direct discrimination*); (ii) *disparate impact* (Barocas and Selbst, 2016; Lipton et al., 2018): outcomes across protected groups differ, thus affecting members of a protected class more than others (*indirect discrimination*). Most approaches to mitigate discrimination are based on the notion of protected/sensitive variables (Caton and Haas, 2020). The protected attributes of an individual include records or visual characteristics that should not impact the model's decision. Common examples are gender, ethnicity, disability, and age. However, the notion of protected variable can encompass any feature of the data that involves people (Barocas et al., 2017).

Recently, there is a great concern in assessing and mitigating biases in order to improve the ethics of the predictions made by automated system, which are increasingly being adopted in our every-day life and can significantly impact human's lives. One of the most well-known example is in the field of judicial system, where recent research have

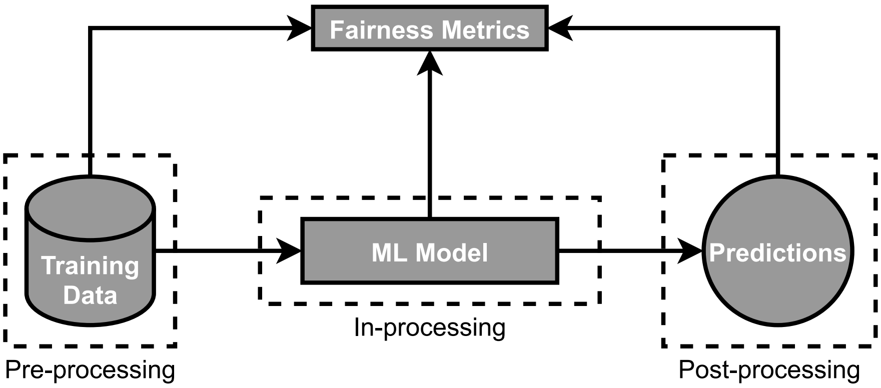


Figure 2.1: Traditional machine learning (ML) pipeline, based on the work of Caton and Haas (2020). This framework has three major steps: training data collection, ML model training, and future outcome prediction. It also includes a stratified view of some fairness initiatives: *pre*, *in*, and *post*-processing, all considering a specific fairness metric.

shown that a tool used by courts in the United States had bias against African-Americans: the tool falsely predicted future criminality twice the rate for African-Americans as it predicted for white people (Angwin et al., 2016; Chouldechova, 2017). Similar findings have been made in other areas, such as hiring applications. For instance, it was recently discovered that Amazon's AI hiring system was discriminating against female candidates, especially for technical positions (Dastin, 2018). In advertising, it was shown that Google's ad-targeting algorithm had proposed more higher-paying executive jobs to men than to women (Datta et al., 2015).

Machine learning algorithms can learn bias from a variety of different sources. Everything from the data used to train it, to the people who are using this tool, and even seemingly unrelated factors can contribute to AI bias (Mehrabi et al., 2021). However, the literature classify all those factors into two potential main causes of unfairness in ML (Mehrabi et al., 2021): those emerging from biases in the data and those emerging from the algorithm. Biases in the data may arise from the dataset itself, based on device measurements, historically biased human decisions, erroneous reports and interpretation, missing data or other reasons. Biases in the algorithm may be related to its objective, which aims at minimizing overall aggregated prediction errors and therefore may benefit majority groups over minorities (Pessach and Shmueli, 2022). Thus, in this last case, minimizing the average loss may result in representation disparity – minority groups contribute less to the training objective and therefore tend to suffer higher loss (or penalty) (Hashimoto et al., 2018).

Much of the related literature that focuses on mitigating ML discrimination algorithms address this either at a technical aspects of bias, or theoretically at a social, legal, and ethical view (Caton and Haas, 2020). A stratified view of some fairness initiatives, as well as a traditional ML pipeline, can be visualized in Figure 2.1. The pipeline is generally composed of three major parts: (1) collecting the training data; (2) optimizing the model with the training data; and (3) predicting future outcomes on a test data. Whilst not all approaches for fair ML fit into this framework, it provides an accepted visual reference (Caton and Haas, 2020; Mehrabi et al., 2021; Dunkelau and Leuschel, 2019). Technical approaches may be applied to the (1) training data, prior to modelling, known as *pre-processing*; at the point of (2) modelling, named as *in-processing*; or at (3) test time, after modelling, called *post-processing*.

Pre-processing approaches argue that the issue is in the data itself, as the distributions of specific sensitive variables may be biased and/or imbalanced. Thus, pre-processing approaches tend to alter the distribution of sensitive variables in the dataset itself. More generally, these approaches perform specific transformations on the data with the aim of removing discriminatory attributes from the training data (Celis and Keswani, 2019). The main idea of this approach is to train a model on this modified version of the dataset. Pre-processing is argued as the most flexible part of the data science pipeline (Caton and Haas, 2020), as it makes no assumptions with respect to the choice of the applied model architecture and training procedure choices.

In the work of Caton and Haas (2020), pre-processing approaches are split into several categories, including adversarial learning, causal methods, relabelling and perturbation, (re)sampling, re-weighting, transformation and variable blinding. Our work fits into the *pre-processing* approach since we add new labels for which different models are then optimized. As far as we know, this is the first time new annotations, specially the ones based on more objective notions (*i.e.*, geometric facial traits) of the attribute are generated to mitigate discriminatory algorithmic behavior. Some approaches (Kamiran and Calders, 2012; Kamishima et al., 2012) have, however, proposed relabelling strategies, which flip or modify the sensitive attribute, or even change the distribution of one or more variables in the training data directly. Usually, relabelling involves the modification of the labels of training or testing data instances so that the proportion of instances are equal across all protected groups. In other words, these approaches often balance the dataset with respect to the target and sensitive attributes, which is not our aim in this work.

Other common approaches in the pre-processing front include balancing the original dataset by augmenting the input images (Ramaswamy et al., 2021; Sattigeri et al., 2019), usually expanding to a great extent the total number of original images contained in the dataset. For instance, in Sattigeri et al. (2019), the authors use a generative model to create data that is similar to a given dataset, but results in a model that is more fair with respect to protected attributes. Another example is the work of Ramaswamy et al. (2021), which modifies the generative model to create new instances by independently altering specific attributes (*e.g.*, by removing glasses) that were found to correlate with a specific sensitive group. They then increase the original dataset size with two times more images.

In contrast with the pre-processing approaches, the in-processing ones argue that the fairness issue may be in the modelling technique (Caton and Haas, 2020). Additionally, as pointed out by Wang et al. (2019), mitigating biases in the dataset itself might be time-consuming and insufficient, since models can still make spurious correlation using other indirect factors. Therefore, this line of research argues that the model can become biased by using dominant and possibly hidden features, or applying a specific objective loss (Mehrabi et al., 2021). Usually, these approaches tackle fairness issues by adding one or more fairness constraints into the model optimization functions towards maximizing performance and fairness. Thus, bias mitigation is usually done via model regularization, which can be implicitly or explicitly applied based on a fairness metric. In the work of Caton and Haas (2020), they propose some sub-categories which are aligned with this view, including adversarial learning, bandits, constraint optimization, regularization and loss reweighting.

Finally, the post-processing approaches claim that the actual output of a model may be unfair to one or more protected variables (Caton and Haas, 2020). Thus, post-processing approaches tend to apply transformations to the model output to improve fairness metrics. This is a flexible approach as it does not need access to the actual algorithms that were used to build the ML model. Moreover, it is applicable for several "black-box" scenarios, where not the entire ML process is known. Some categories for post-processing approaches include calibration, constraint optimization, thresholding and transformation (Caton and Haas, 2020).

2.1.1 Ensembles

Ensemble methods is a machine learning technique that combines several base models in order to produce an optimal predictor. Based on the intuition that diversity of decisions is an essential element to achieving fair outcomes, recent work have been focused on studying the impact of ensemble models as a method for reducing unfair behavior. The work of Grgić-Hlača et al. (2017) was one of the pioneers in this field. They studied the fairness properties of ensembles when a classifier is randomly selected to make the prediction. They presented settings in which an ensemble of unfair classifiers can be fair, and settings where an ensemble of classifiers can provide better fairness-accuracy trade-offs than individual predictors.

While Grgić-Hlača et al. (2017) points out the potential of using ensemble to improve model fairness, their work lack empirical evidence. Thus, Bhaskaruni et al. (2019) proposed a new ensemble learning strategy for fair learning that adopts the AdaBoost framework. However, unlike the original AdaBoost that upweights mispredicted instances, their method upweights unfairly predicted instances which are identified by an adaptation of the k-Nearest Neighbors (kNN) method presented in Luong et al. (2011). They considered the case where the generated classifiers are dependent, *i.e.*, one is built on top of the other. Their technique empirically demonstrated how ensembles have useful propertied for a specific fairness metric named statistical parity.

One drawback of their approach is that the method showed a significant accuracy drop. Kenfack et al. (2021) then proposed another weighting technique, in which the weights are assigned proportionally to the predictors' performance in terms of accuracy and fairness. Their work focuses on the fairness performances of ensemble models built using independent classifiers and without any fairness constraints. They demonstrated that their proposed technique allows to reduce biases of the ensemble while maintaining certain accuracy compared to uniformly weighted classifiers and individual classifiers.

More recently, the work of Feffer et al. (2022) studies whether combining bias mitigation techniques and ensembles can help with fairness in several different scenarios. They conduct an extensive set of experiments, combining 10 bias mitigators, 4 ensembles and 13 datasets, and demonstrated that ensembles can often improve stability of accuracy as well as group fairness metrics. However, they focused on tabular data, and their results showed that the best configuration of bias mitigation strategy and ensemble depends on several factors, including dataset characteristics, learning objectives, and worldviews (Feffer et al., 2022).

Thus, our work is inspired by the existing literature on ensembles for reducing unfair outcomes. However, differently from previous approaches that mainly present new weighting techniques, we propose combining image classification models using a simple weighted average. Our novelty is the fact that each of the base models that compose the final ensemble is trained on a different perspective of the task at hand. These perspectives are treated in the label level, and split into subjective and objective notions of the attribute, as we describe in more detail in Section 3.2.

2.1.2 Evaluation Metrics

Many different metrics (Barocas and Selbst, 2016; Berk et al., 2018; Chouldechova, 2017; Hardt et al., 2016; Kleinberg et al., 2016) have been proposed to mathematically measure and represent bias, fairness and/or discrimination of machine learning models, especially for binary classification. Although the literature has defined a myriad of notions to quantify fairness notions, each one has different aspects of what is be considered “fair”. This is mainly a consequence of having various different interpretations of what means for an algorithm to be considered “fair”. The fairness notions may be subdivided into two different views: *individual fairness*, which argues that similar individuals (inputs) should yield similar predictions (outputs); and *group fairness*, which defends that instances should be grouped according to a specific sensitive attribute and that the algorithm should behave similarly across all groups.

Most quantitative definitions and measures of fairness are centered around three fundamental aspects of a (binary) classifier (Caton and Haas, 2020): First, the sensitive variable S that defines the groups (for group fairness) or individual attributes (for individual fairness) for which we want to measure fairness. Second, the target variable Y, which has, in binary classification scenario, two possible values ($Y = 0$ or $Y = 1$). Third, the classification score P, which represents the predicted score (usually defined as probabilities, limited within the interval $[0, 1]$) that a classifier outputs for each instance.

Regarding individual fairness metrics, the definition of the similarity might depend on a specific task (Dwork et al., 2012), and it may be generally described as follows:

$$|P[\hat{Y}_i = y|X_i, S_i] - P[\hat{Y}_j = y|X_j, S_j]| \leq \epsilon; \text{ } \mathit{if} \text{ } d(i, j) \approx 0, \tag{2.1}$$

where $\hat{Y}_i$ is the model's output prediction for individual i, S_i refers to the sensitive attributes of individual i, and X_i refers to his/her associated features (inputs). ϵ is the minimum amount of unfairness accepted by the system. The distance metric $d(i, j)$ between individuals i and j can be defined depending on the domain, *e.g.*, KL Divergence or L1-norm (Biega et al., 2018), and as noted by Pessach and Shmueli (2022) choosing this distance metric is not a trivial task.

For group fairness, different kinds of measurements have been proposed. For example, *demographic parity*, also known as *statistical parity* and *Calders-Verwer discrimination score* ($\Delta Disc$), requires that the average of positive predictions be similar across different groups (Calders and Verwer, 2010). Formally, it is computed as:

$$|P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S = 0]| \leq \epsilon, \tag{2.2}$$

where P is the output probabilities, $\hat{Y}$ are the model predictions, where $\hat{Y} = 1$ represents positive predictions (*e.g.*, acceptance of a specific condition, such as for a job), and S represents the protected attributes (*e.g.*, race and gender), where, usually, $S = 1$ are the instances of the privileged group and $S = 0$ are the ones of the unprivileged group. A lower value indicates more similar acceptance rates and thus more fairness. Intuitively, this metric tries to ensure that the positive prediction is assigned to the two sensitive groups at a similar rate. It is also independent of the ground-truth labels (Du et al., 2020), which has the advantage of avoiding the mislabeling problem, *i.e.*, incorrectly labeling a set of instances, and it circumvents the cost of gathering high-quality annotations. However, one limitation with this view is that different groups could have very different label distribution for Y (Beutel et al., 2019).

Hardt et al. (2016) addresses this by proposing *equalized odds*, which computes the difference between the false positive rates (FPR), and the difference between the true positive rates (TPR) of the two groups. It can be formalized as follows:

$$|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S = 0, Y = 0]| \leq \epsilon, \tag{2.3}$$

$$|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S = 0, Y = 1]| \leq \epsilon, \tag{2.4}$$

where Equation 2.3 requires the absolute difference in the FPR of two groups to be bounded by ϵ, and Equation 2.4 represents the absolute difference in the TPR of two groups bounded by ϵ. Smaller difference between groups indicates better fairness. Intuitively, this metric proposes that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members (Verma and Rubin, 2018).

A more specific metric based on equalized odds is *equality of opportunity* (ΔEoO) (Hardt et al., 2016), which some works (Garg et al., 2020; Zhang et al., 2018a) define as the absolute difference between true positive rates (TPR) across sensitive groups, while other works (Castelnovo et al., 2022; Sattigeri et al., 2019) define as the absolute difference between false negative rates (FNR) across the groups. We follow the latter definition since previous related literature on the task we are tackling in this work uses the latter definition. It can be formalized as:

$$\Delta EoO = |P[\hat{Y} = 0|S = 1, Y = 1] - P[\hat{Y} = 0|S = 0, Y = 1]|. \tag{2.5}$$

Even though many metrics have been proposed to mitigate discriminatory behavior, some work (Corbett-Davies et al., 2017; Berk et al., 2018; Chouldechova, 2017; Friedler et al., 2021; Corbett-Davies and Goel, 2018) have shown that it is not possible to simultaneously satisfy several metrics (notions) of fairness. Moreover, the literature discusses an existent trade-off between accuracy and fairness – as one increase fairness, it is possible to compromise accuracy (Corbett-Davies et al., 2017; Kleinberg et al., 2016). Overall, the goal of mitigating unfairness in algorithms is to achieve a model that allows for higher fairness without significantly compromising the accuracy or other alternative notions of utility.

To measure the discriminatory behavior of the models, we use different metrics according to the related work of the two facial analysis tasks explored in this work. For the attractiveness classification (Sattigeri et al., 2019; Ramaswamy et al., 2021), we use the metric of Equality of Opportunity (ΔEoO), defined in Equation 2.5. For the FER (Chen and

Joo, 2021), we use the Calders-Verwer discrimination score (Calders and Verwer, 2010) ($\Delta Disc$), defined in Equation 2.2.

2.2 Attractiveness

Attractiveness has a wide variety of impacts on peoples' lives. The pursuit for beauty has encouraged trillion-dollar cosmetics, aesthetics, and fitness industries, each one promising a more attractive, youthful, and physically fitter version for each individual (Cutler, 2021). Beauty also seems to be an important aspect of human social interactions and matching behaviors, in which more attractive people appear to benefit from higher long-term socioeconomic status and are even perceived by peers as "better" people (Cutler, 2021). In fact, attractive individuals are perceived to own a variety of positive personality attributes (Little, 2014). They are considered to be friendlier, more intelligent, more interesting, and more socially competent (Hönn and Göz, 2007).

There have been a large number of studies examining attractiveness stereotypes. In the work of Cash and Kilcullen (1985), hiring preferences were shown for attractive over unattractive applicants. Attractiveness may also influence judgements about the gravity of committed crimes (Sigall and Ostrove, 1975). Moreover, magazines, social media, and television are filled with attractive faces, which may in fact reinforce the hypothesis that the media exploits universal beauty standards. This is especially true in online social media platforms (Cutler, 2021), which have influenced beauty standards through instantaneous editing features, filtering, and cropping, allowing people to become the ideal version of themselves. As some studies demonstrate (Sherlock and Wagstaff, 2019; Verrastro et al., 2020) such tools have a great impact on individuals and society in general.

Even though attractiveness has a considerable influence over our lives, which characteristics make a particular face attractive is imperfectly defined. It has been shown, however, that there is a very high agreement between groups of raters belonging to the same culture and even across cultures (Cunningham et al., 1995; Langlois et al., 2000). Thus, if different people can agree on which faces are attractive and which faces are not attractive when judging faces of different ethnic backgrounds, then this indicates that people all around the globe use similar features or criteria when making up their judgments.

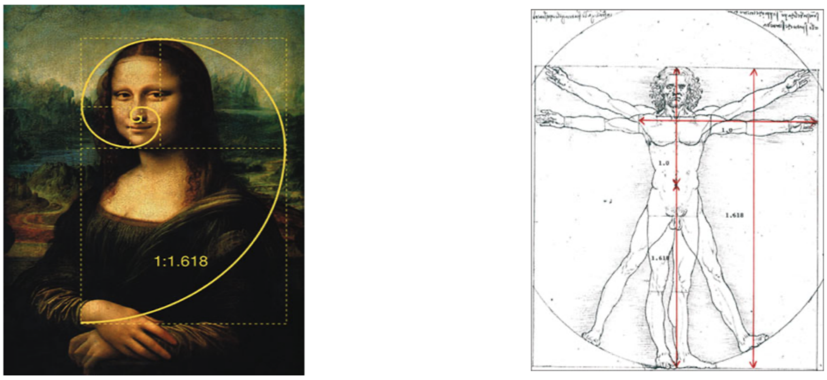
Further evidence shows that infants prefer to look at faces that are rated by adults more highly for attractiveness than at those faces rated lower (Samuels et al., 1994; Langlois et al., 1987). This supports the theory that there exists something innate about attractiveness, which is not influenced by culture nor background. Thus, both early developmental and cross-cultural agreement on attractiveness are initial evidence against the notion that attractiveness ideals are slowly absorbed by growing up within a particular culture, and this suggests that there is indeed a global consensus on which features are considered attractive, and which features are not.

Based on the evidence that the notion of attractiveness is universal, some automated systems for rating facial attractiveness were proposed. The earliest classifiers for this task are based on traditional machine learning methods. As well as for other computer vision tasks, deep learning brought great performance improvements for facial beauty assessment (Gan et al., 2014; Gray et al., 2010). However, as recently demonstrated (Ramaswamy et al., 2021; Sattigeri et al., 2019), these approaches were shown to discriminate against certain groups.

Some pre-processing approaches were proposed to mitigate such behavior. For instance, in the work of Sattigeri et al. (2019), the authors modify the training loss of a generative model to create a synthetic dataset that is similar to a given dataset, but results in a model that is more fair with respect to protected attributes for either demographic parity or equality of opportunity. Another example is the work of Ramaswamy et al. (2021), which balanced the training data with respect to the protected attributes using a generative model. They modified this model to create new instances by independently altering specific attributes (*e.g.*, removing glasses) to de-correlate the target label and the sensitive attribute. Both works expand the training dataset from two to three times its original size.

In addition to facial features, shape, and form, people judge human faces using several other attributes such as expressions, average (composite of) faces, thirds and fifths theory (Gunes, 2011; Kagian et al., 2007). The multiple fitness model (Cunningham et al., 1995) suggests that there is no single feature or dimension that determines attractiveness. Rather, attractiveness is determined by a combination of several features, which individually represent different aspects of a persons' face. This theory still supports that some facial qualities are perceived as universally (physically) attractive.

In this work, we build upon this idea by proposing a method that combines different learners, each trained on a different perspective of the notion of attractiveness. We hypothesize that each model possibly captures different features to compose its final prediction. To extract the so-called *objective* annotations of the attractiveness based on facial traits, we follow the work of Schmid et al. (2008) and use three predictors that have been proposed in literature: Golden Ratios (Section 2.2.1), Neoclassical Canons (Section 2.2.2), and Facial Symmetry (Section 2.2.3). All of them were empirically shown to correlate with human attractiveness ratings (Schmid et al., 2008). We describe these notions in the following sections.



(a) Mona Lisa, by Leonardo da Vinci. (b) Human figure, by Leonardo da Vinci.

Figure 2.2: Popular examples of the use of the Golden Ratio theory (Zhang et al., 2016).

2.2.1 Golden Ratio

The Golden Ratio theory, also known as the divine proportion, can be seen in several different fields, such as arts, architecture, and flowers (Zhang et al., 2016). One famous example is the painting of Mona Lisa, which embodies the golden ratio, as can be visualized in the left side of Figure 2.2. Another well-known example that uses this proportion is Leonardo Da Vinci's drawing of the human body, which can be seen in the right side of Figure 2.2. This theory defines that faces that have features with ratios close to the golden ratio proportion are perceived as more attractive (Gunes, 2011). The golden ratio is approximately the ratio of 1.618 to 1 (Gunes, 2011). Schmid et al. (2008) demonstrated that six facial ratios are predictors of attractiveness. These ratios are defined in Table 2.1, where *x* and *y* refers to the *x*-coordinate or *y*-coordinate of the points, and the numbers indicate which points of Figure 2.3 were used in their study to calculate the ratio.

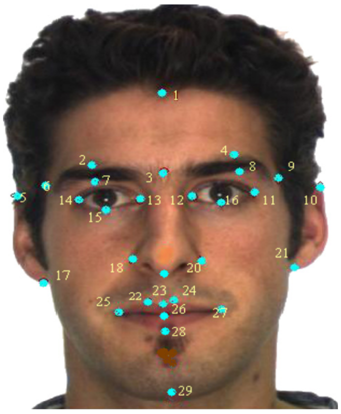


Figure 2.3: Feature points of a facial image from Schmid et al. (2008).

Description	Numerator Points	Denominator Points
Mouth width to interocular distance	$x\_25 - x\_27$	$x\_12 - x\_13$
Lips–chin distance to interocular distance	$y\_23 - y\_29$	$x\_12 - x\_13$
Lips–chin distance to nose width	$y\_23 - y\_29$	$x\_18 - x\_20$
Length of face to width of face	$y\_1 - y\_29$	$x\_17 - x\_21$
Mouth width to nose width	$x\_25 - x\_27$	$x\_18 - x\_20$

Table 2.1: Golden ratio definitions shown to correlate with attractiveness ratings by humans (Schmid et al., 2008).

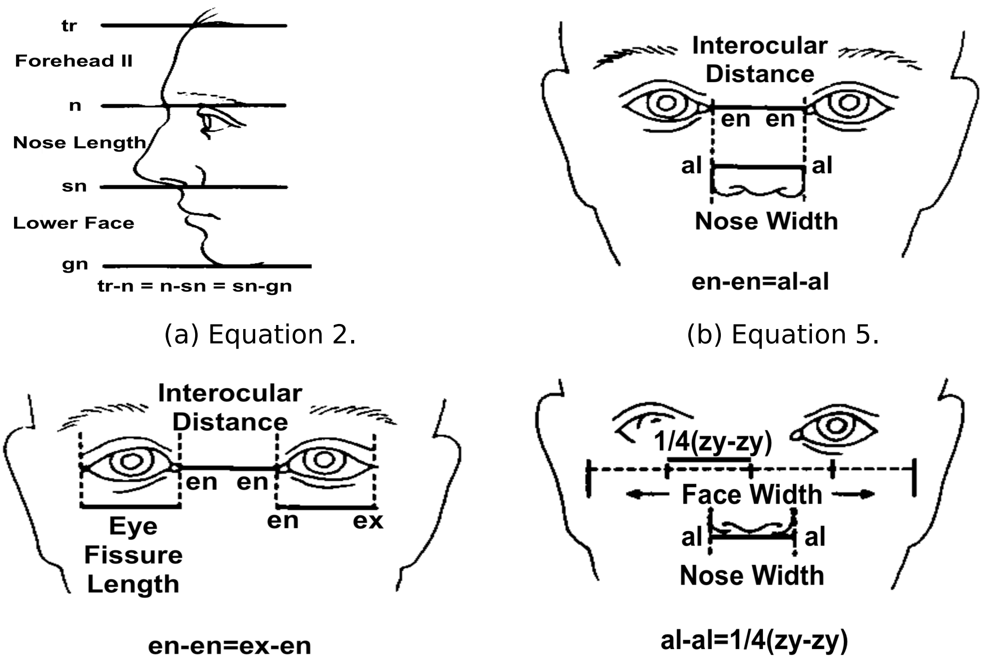
2.2.2 Neoclassical Canons

Neoclassical canons were proposed by artists in the renaissance period as guides for drawing beautiful faces (Farkas et al., 1985). The basic idea behind this definition is that the proportion of an attractive face should follow some predefined ratios. Farkas et al. (1985) summarize these principles into nine neoclassical canons and their variations. Four of the canons use vertical measurements, four use horizontal measurements, and one uses angles of inclination. Only six of them can be tested from the frontal views of the facial images. However, Schmid et al. (2008) show that only five of the six frontal ratios have a significant correlation with attractiveness. We use all of the human-correlated ratios, except one which uses ear points, since we only extract facial landmarks. All of them are defined in Table 2.2 and illustrated in Figure 2.4.

Description	Equation Number
Forehead height = Nose Length = Lower Face Height	2
Interocular Distance = Nose Width	5
Interocular Distance = Right of Left Eye Fissure Width	6
Face Width = 4× Nose Width	8

Table 2.2: Neoclassical canons definitions shown to correlate with attractiveness Schmid et al. (2008). Equation numbers are from Farkas et al. (1985).

Additionally, we follow the procedure defined in Schmid et al. (2008) and compute the coefficient of variation instead of the individual distances, since some measures use more than two features (*e.g.*, Equation Number 2 in Table 2.2). The coefficient of variation (cv) computes the ratio of the standard deviation (σ) of the distances to the mean (μ) of the distances ($cv = \frac{\sigma}{\mu}$). For instance, to compute the final distance in Equation Number 2, from Table 2.2, would require pair-wise comparisons of three distances. The use of the coefficient of variation allows us to incorporate all three distances into one value while adjusting for the size of the face (dividing by the mean). The larger the coefficient of



(a) Equation 2.

(b) Equation 5.

(c) Equation 6.

(d) Equation 8.

Figure 2.4: Neoclassical canon equations from Farkas et al. (1985), shown by Schmid et al. (2008) to correlate with human attractiveness ratings.

variation, the more the face differs from the canon. A value of zero indicates there is no variation in the distance, *i.e.*, they are equal.

2.2.3 Symmetry

Facial symmetry is considered an important factor for attractiveness (Rhodes, 2006). Symmetry has many different definitions (Schmid et al., 2008; Rhodes, 2006; Kowner, 1996; Perrett et al., 1999), but it generally refers to the extent that one half of an image (*e.g.*, face) is the same as the other half. In our work, we follow the definition presented in Schmid et al. (2008), which define the axis of symmetry to be located vertically at the middle of the face. To characterize this line they fit the least squares regression line through the seven points measured along the middle face (Points 1, 3, 19, 23, 26, 28, 29 depicted in Figure 2.3). In this work, the following facial attribute pairs (left and right) are used (Schmid et al., 2008):

- Eyebrows (Points 2 and 4; Points 7 and 8)
- Eyes (Points 11 and 14; Points 12 and 13; Points 15 and 16)
- Nose (Points 18 and 20)
- Lips (Points 22 and 24; Points 25 and 27)

- Face (Points 6 and 9)

We used all the symmetry ratios shown to correlate with human attractiveness ratings in Schmid et al. (2008), except the one which uses ear points since we only extract facial landmarks. To compute the symmetry of a face, the authors propose firstly computing the symmetry of the individual features, *i.e.*, the distance between the points described above, and then averaging them all to obtain a single value for each face. To compute the distance metric, we use the function d_{p_L,p_R}, where d is defined as the euclidean distance between two coordinates for the left (p_L) and the right (p_R) sides. The euclidean distance is further defined as:

$$d_{p,m} = \sqrt{(p_x - p_{m_x})^2 + (p_y - p_{m_y})^2}, \tag{2.6}$$

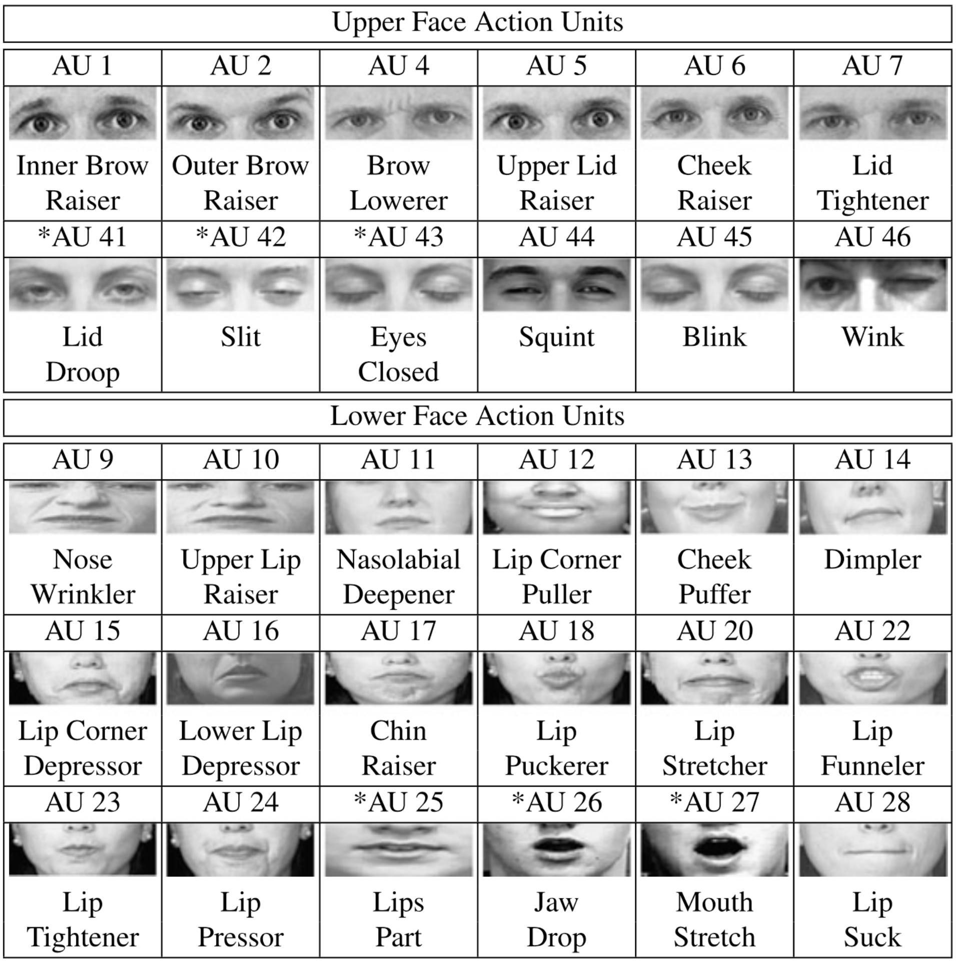
where p represents a point in the face, and p_m represents a point in the middle of the face, both in terms of x, y axis. A value of zero in the distance implies ideal symmetry; the greater the absolute value the less symmetric the face is.

2.3 Facial Expression Recognition

Facial expression is one of the most powerful and instinctive means of communication for human beings. A facial expression can be considered as a visible manifestation of an inner state of mind and hence gives idea about intention, interest, and psychology of a person. The study of facial expression analysis dates back to the Aristotelian era (Bettadapura, 2012). One of the important works on facial expression analysis, that had and still has a huge impact on modern days is the work of Darwin (2014). He established the means of expressions in humans and animals, and categorized them into different groups.

Another important milestone in the study of facial expressions is the work done by Ekman and Friesen (1978), who developed the Facial Action Coding System (FACS). FACS is a muscle-based approach that identifies the facial muscles that individually, or in groups, cause changes in facial behaviors. These changes in the face and the underlying muscle(s) that caused these changes are called Action Units (AUs). FACS is composed of a total of 44 AUs, and each one can be described in two ways: (1) presence: if the AU is visible (*i.e.*, active) in the face. (2) intensity: how intense is the AU (minimal to maximal) on, usually, a 5-point ordinal scale. Some AUs can be visualized in Figure 2.5. The work of Ekman and Friesen (1978) is of significant importance and still has a large influence on the development of facial expression recognition (FER) systems.

With some exceptions (*e.g.*, Tian et al. (2001)), most modern automated FER systems are treated as a classification problem, and they attempt to recognize a small set of emotional expressions (classes) as shown in Figure 2.6 (*i.e.*, disgust, fear, happiness, sur-



Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink

Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 2.5: Facial Action Units (AUs). The AUs with “$*$” indicate that the criteria have changed for this AU, that is, AU 25, 26, and 27 are now coded according to criteria of intensity (25A-E), and AU 41, 42, and 43 are now coded according to criteria of intensity (Tian et al., 2011).

prise, sadness, anger), besides the neutral expression (no emotion, thus no active AU). As in other related fields, deep learning has improved the performance of FER systems, and some works (Xu et al., 2020; Chen and Joo, 2021) have recently focused on understanding and mitigating biases in such systems. For instance, Li and Deng (2020) observed that disgust, anger, fear, and surprise are usually underrepresented classes in datasets, thus being harder to learn compared to the classes that have the majority of samples in the dataset.

Another study conducted by Rhue (2018) provides evidence that some real-world applications of facial recognition interpret emotions differently based on the person’s race. For instance, Face++ consistently interprets black players as angrier than white players, even controlling for their degree of smiling. Another example is the Microsoft’s system, which registers contempt instead of anger, and it interprets black players as more contemptuous when their facial expression is ambiguous.

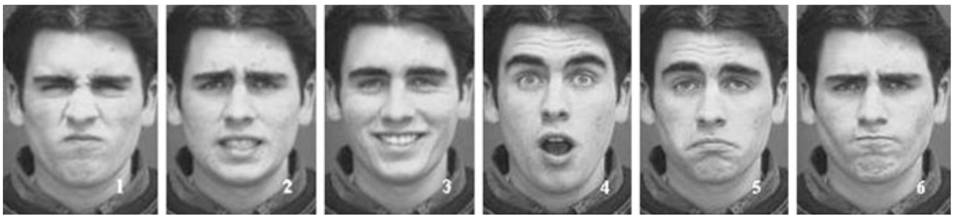


Figure 2.6: Posed facial expression (images from database Kanade et al. (2000)). 1: disgust; 2: fear; 3: happiness; 4: surprise; 5: sadness; 6: anger.

Moreover, some works have shown slight differences in perception regarding some expressions in female and male faces. For instance, women were shown to be generally seen as happier than men (Steephen et al., 2018). Becker et al. (2007) demonstrated that people are faster and more accurate at detecting angry expressions on male faces and happy expressions on female faces. Denton et al. (2019) found that a smiling classifier trained on the CelebA dataset (Liu et al., 2015) is more likely to predict ‘smiling’ when eliminating a person’s beard or applying makeup or lipstick to the image while keeping everything else unmodified.

Among the common approaches that try to reduce unfair behaviors are the ones that are generic to a set of facial classification tasks. For instance, the work of Alvi et al. (2018) proposes a debiasing approach based on domain adaptation. Specifically, they create a network with several output branches, where the primary branch has a single classification loss, which assesses the ability of the network to accurately distinguish between classes of the primary task, and several secondary branches which have two losses: a classification loss and a confusion loss. These losses are used to, in turn, assess the amount of spurious information in the feature representation and then remove it. Another work among this line is the one from Wang et al. (2020), which studied the mitigation strategies of data balancing, fairness through blindness, and fairness through awareness, and demonstrated that fairness through awareness provided the best results for smiling/not-smiling classification on the CelebA dataset (Liu et al., 2015).

Nonetheless, other approaches have been designed specifically to mitigate biases in the FER systems. In the work of Xu et al. (2020), for instance, they studied the effect of methods that use confusion loss for mitigating biases in the RAF-DB dataset (Li et al., 2017). Specifically, they analyzed an ‘attribute-aware’ network, where the classification layer of the network receives a representation of the attributes and can explicitly use this information to model the effect of biases, and a ‘disentangled approach’, where the network has several parallel classification branches, one for the main task and others for each sensitive attribute, and a confusion loss is used on top of them so that the sensitive attributes is not predictable from the main classification head. In other words, the confusion loss aims at de-correlating the target and sensitive attributes.

Recently, in the FER field, the idea of using the relationship among multiple labels has been explored (Chen and Joo, 2021). In the context of objective labels to mitigate fairness issues, Chen and Joo (2021) leads the initiative by proposing an in-processing approach that incorporates the triplet loss to embed the dependency between Action Units (AUs) and expression categories. The main idea behind the method is to encourage the model to treat two samples in a similar way if the AUs that compose the facial expression are similar, even if their gender expressions and annotated labels are different. Differently than their work, we propose to use AUs in the pre-processing step, as a novel and more objective labeling strategy.

2.4 Explainability

Advances in machine learning and deep learning have had a profound impact on many domains. Recently, researchers have begun to explore how these approaches can be used in high-stake domains such as healthcare, criminal justice system, finance, and military decision making (Chakraborty et al., 2017; Goodfellow et al., 2016). As the importance of the decisions aided using ML increases, it becomes increasingly important for users to be able to suitably weight the assistance provided by such systems. A fundamental property is *explainability* – ML models should provide the relevant parts of the machine representations in an understandable format for humans. In other words, explainability refers to the type and completeness of the output given when a model is queried for the reasoning behind its decision (Chakraborty et al., 2017).

Since the approval of the European Union's new General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019), the use of automated individual decision-making has been restricted, and the new directives focused on the protection of sensitive data of persons, such as their age, gender, ancestry, name, or place of residence, for instance. The GDPR also imposed a data quality requirement in the AI area since the quality of the training datasets has a great impact on the outcome (Jain et al., 2020). The GDPR gives any citizen the "right to explanation" of an algorithmic decision made about them (Buhrmester et al., 2021). The explanation has to be transmitted in a precise, transparent, understandable, and easily accessible form and in a clear and simple language.

Some ML models are inherently interpretable since its conception, *e.g.*, linear models or decision trees (Rudin, 2019). However, most ML models, especially in the deep learning (DL) domain, are referred to as *black box*, and usually require additional mathematical frameworks to explain their behavior. Based on the scope and the purpose of the explanation (Nielsen et al., 2022), a user can employ a local or a global explainability method. Global explainability methods attempt to explain the overall decision-making process of the model, *i.e.*, how the inputs are transformed into the output decisions at the model level. In contrast, local explainability methods attempt to explain individual decisions, *i.e.*, what features of a specific input (*e.g.*, pixels of an image) may have influenced the model's decision.

Two broad categories of local explainability exist, namely, methods based on *feature perturbation* and others based on *gradient information* (Nielsen et al., 2022). The former class of methods records the effect of masking or altering the input features onto the network's performance. This requires multiple passes through the network to determine the importance of each input pixel, making perturbation-based attribution computationally intensive. In the latter case, the gradients of the output (logits or softmax probabilities) with respect to the extracted features or the input are calculated via backpropagation and are used to estimate attribution scores. The magnitudes of gradients show the importance of input to output scores.

In our work, to understand which features contributes mostly to the models' final decisions, we use the saliency method (Simonyan et al., 2014), a local gradient-based method that generates heatmaps to point at where an AI model is attending to as it makes a specific decision or prediction. The heatmaps have been treated as explanations, and they recently became a popular alternative for explaining deep neural network behavior (Tomsett et al., 2020). Specifically, we use the technique developed by Simonyan et al. (2014). It uses Taylor series, based on partial derivatives to display input sensitivities in images. Their saliency maps were shown to represent the first-order approximation of the attributions.

2.5 Context of this work in the Literature

Our work aims to propose a new method for mitigating unfair behavior of ML systems by incorporating more diverse and objective labels to the training procedure of the models that compose the final ensemble. For combining the models, we take inspiration from previous work (Grgić-Hlača et al., 2017), described in Section 3.2.3, which showed that ensembles can be a useful technique for reducing biases in ML systems. However, previous work (Bhaskaruni et al., 2019; Kenfack et al., 2021; Feffer et al., 2022) that exhibited empirical evidence of such fact focused mainly on proposing new weighting mechanisms that deal with the accuracy and/or fairness performance of the models that compose the final ensemble. We do not design a new weighting mechanism, instead, we make use of a simple weighted average, and center our attention to the diversity of labels used to train the individual models.

In our work, we focus our effort on two main tasks: (1) attractiveness classification, and (2) facial emotion recognition. For the first task, some works (Sattigeri et al., 2019; Ramaswamy et al., 2021), which we describe in Section 2.2, have recently addressed the fairness problem by proposing pre-processing approaches that generate synthetically modified images to balance the training instances with respect to the sensitive attributes. Most of these approaches use Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). In some sense, this is similar to our approach, in which it tries to incorporate more diversity to the data. However, both approaches significantly increase the computational cost for (1) creating synthetic images, and (2) training the model with the augmented dataset. Moreover, the generative models may create non-realistic images, and even include their own biases into the dataset. Instead of creating new instances for the dataset, we propose annotating the original instances with new labels, *i.e.*, we add new perspectives on the notions of each task.

The second task of facial expression recognition has not yet received wide attention (Xu et al., 2020; Chen and Joo, 2021; Rhue, 2018). All of the past approaches that we are aware of, which we describe in Section 2.3, have proposed in-processing methods for mitigating biases in this task. Specifically, Xu et al. (2020) conducted a study on confusion loss for two datasets, named RAF-DB (Li et al., 2017; Li and Deng, 2019) and CelebA (attribute smiling/not smiling) (Liu et al., 2015). Chen and Joo (2021) conducted a study on annotation bias, and found that systematic biases exist in many facial expression datasets, especially the ones collected in-the-wild. They propose to incorporate the triplet loss into the objective function to embed the dependency between AUs and expression categories. This is the first work that proposes using some form of objective annotation procedure, however, they use it in the in-processing step. In contrast, we propose generating objective labels to boost diversity in the decision-making process.

Thus, to the best of our knowledge, our work is a first step towards proposing a labeling strategy, which incorporates both objective and subjective labels, as a pre-processing method for augmenting the variety of individual models in the final ML system.

3. DATA AND PROPOSED METHOD

In this chapter, we initially present preliminary information regarding this study (Section 3.1). Specifically, we start by clarifying the notation (Section 3.1.1) used throughout this work, as well as evaluation metrics (Section 3.1.2) and datasets (Section 3.1.3) for each task (*i.e.*, attractiveness classification and facial expression recognition). We then provide a detailed description of our proposed method, which consists of three main steps: (1) data annotation based on different mathematical notions of attractiveness and facial expression; (2) next, we train one machine learning model for each of the mathematical concepts, and one model with the original human-based annotations; and (3) finally, we aggregate all the models into an ensemble framework. We weight the combination of individual models, each trained on specific attractiveness/facial expression notion. Our main hypothesis is that by combining the objective (geometrically-based), and the biased and subjective (human-based) notions we can effectively reduce the effect of discrimination on the system. Our goal is to create a diverse set of decision-making algorithms that when combined can produce a fairer system.

3.1 Preliminary

In this section, we describe preliminary information that is useful for the deep understanding of our work. We start by formalizing the notation applied in this work. We then introduce the evaluation metrics used to assess our models. Finally, we describe the datasets used to train and evaluate our models.

3.1.1 Notation

We define a dataset $\mathcal{D}$ consisting of $\mathcal{D} = (S, X, Y)$, where S represents sensitive attributes such as gender, age, and skin color, X represents input features, and Y represents the annotated labels. We suppose there are N samples in total and we use S_i, X_i, Y_i to represent the features of the i-th sample. To simplify, we suppose the sensitive attribute S and the outcome Y are binary, which means $Y, S = \{0, 1\}$. In order to train a machine learning model on dataset $\mathcal{D}$, one must optimize parameters θ of function $f_\theta : X \rightarrow \mathbb{R}^{|Y|}$ to produce accurate predictions. In this work, we minimize the standard cross-entropy (CE) loss ($\mathcal{L}$) over dataset $\mathcal{D}$ to provide the correct prediction $\hat{Y}$. A fair machine learning problem is to design a fair predictor $\hat{Y}$ with parameters $\theta : X \times S \rightarrow Y$, which maximizes the likelihood $P(Y, X, S|\theta)$ while satisfying some specific fair constraints that we introduce in the next section.

3.1.2 Evaluation Metrics

In this section, we describe the evaluation metrics we used in this work. Evaluation metrics are a part of every machine learning pipeline, and their goal is to monitor and measure the performance of a given model. In this work, following previous work (Sattigeri et al., 2019; Ramaswamy et al., 2021; Chen and Joo, 2021) we measure the model's usefulness by its overall accuracy, which is is defined as the number of correct predictions divided by the total number of predictions. In some analysis we also provide the accuracy stratified by the sensitive group, *i.e.* accuracy regarding the groups $S = 0$ or $S = 1$. This is calculated by gathering the number of correct predictions by the total number of predictions for instances pertaining to that particular sensitive group according to the annotated labels.

To measure the discriminatory behavior, we use different metrics according to the related work of the two tasks explored in this work. For the *attractiveness* classification (Sattigeri et al., 2019; Ramaswamy et al., 2021), we use the metric of Equality of Opportunity (ΔEoO), which is defined (Sattigeri et al., 2019) as the difference of conditional false negative rates across groups. Formally, it is defined as in Equation 2.5. For the *facial expression* recognition (Chen and Joo, 2021), we use the Calders-Verwer discrimination score (Calders and Verwer, 2010) ($\Delta Disc$), which is defined as the difference between conditional probabilities of advantageous decisions for non-protected and protected members. Formally, it is defined as in Equation 2.2. For more information on the metrics, we refer to Section 3.1.2.

3.1.3 Dataset

In this section, we describe the datasets used for each of the tasks explored in this work. We highlight the fact that we chose the datasets, and used the training and test splits, according to previous research on the fairness field. We start by introducing the dataset used for the attractiveness task (Section 3.1.3), and then describe the datasets used for the facial expression recognition (Section 3.1.3).

Attractiveness

In this work, we use the CelebA dataset (Liu et al., 2015), which is a well-known benchmark in the ML community (Shen and Liu, 2017; Heusel et al., 2017; Choi et al., 2018). We purposefully selected a dataset containing celebrities since previous work already claimed that some aspects of famous people might influence the way people rate attractiveness (Thwaites et al., 2012). For instance, it is suggested by Thwaites et al. (2012)

that the humor and personality associated with a specific character make the celebrity attractive. Additionally, previous work (Sattigeri et al., 2019; Quadrianto et al., 2019; Ramaswamy et al., 2021) already studied and demonstrated fairness issues regarding the attractive feature in this dataset.

This dataset contains celebrity faces with 40 binary face attributes, corresponding to the presence (1) or absence (-1, which in this work we treat as 0) of each facial trait. Each attribute was annotated by a "professional labeling company", further described by Böhlen et al. (2017) as "a group of 50 paid male and female participants, aged 20 and 30, and recruited from mainland China during a 3 month development phase" (Sattigeri et al., 2019). It contains a total of $202,599$ images downloaded from the internet, cropped and resized to 128×128 pixels (Denton et al., 2019). For the attractiveness classification task, we consider three different sets of protected attributes: gender expression and age, which are given in the dataset, and a binary skin tone label that we annotate using the Fitzpatrick skin type scale (Sattigeri et al., 2019). The procedure we followed for annotating the skin color is describe in Section 3.2.1.

Facial Expression

Collecting a large amount of labeled training data that include as many variations of the populations and environments as possible is important for the design of a deep expression recognition systems (Li and Deng, 2020). However, as in other fields of deep learning, this is rarely the case. In our experiments, we follow the procedures and splits used in Chen and Joo (2021). In their work, one dataset was used for training on the happiness attribute, and another one was used for training on the anger attribute. The one used for the former attribute is named Expression in-the-Wild Database (ExpW) (Zhang et al., 2015, 2018b), and the one for the latter is called AffectNet (Mollahosseini et al., 2017). Since we analyse two main facial expressions (*i.e.*, anger and happiness), for the this classification task, we consider only one protected attribute, which is gender expression. We leave other sensitive attributes for future work.

The ExpW dataset (Zhang et al., 2015, 2018b) contains 91,793 faces downloaded using Google image search. Each of the facial images were manually annotated as one of the seven basic expression categories. Nonface images were removed in the annotation process. Since the ExpW dataset does not contain labels regarding sensitive attributes, in the work of Chen and Joo (2021) they annotate all instances with their corresponding gender expressions according to a ResNet-34 (He et al., 2016) gender classifier trained on FairFace dataset (Karkkainen and Joo, 2021). We contacted the authors, who provided access to all the code and data necessary for the experiments.

The AffectNet dataset (Mollahosseini et al., 2017), used in this work for training the models on the anger attribute, is considered one of the largest FER dataset (Li and

Deng, 2020). It contains more than one million images, of which $\approx$ 420,000 of them have manually annotated labels of the presence of seven discrete facial expressions (categorical model) and the intensity of valence and arousal (dimensional model). The remaining ($\approx$ 550,000) were automatically annotated using ResNext Neural Network trained on all manually annotated training set samples with an average accuracy of 65%. All the images were obtained by querying three different search engines using 1,250 emotion related keywords in six different languages.

A third dataset was used for the evaluation part, which is named Chicago Faces Database (CFD) (Ma et al., 2015). The main CFD set consists of images of 597 unique individuals, and it includes self-identified Asian, Black, Latino, and White female and male faces, recruited in the United States. In the work of Chen and Joo (2021), they construct a balanced subset of the test dataset for each expression. First, they train a “naive classifier” (Chen and Joo, 2021) for the attribute at hand, using either the original ExpW dataset (for the happiness attribute) or the original Automatic AffectNet (for the anger attribute). Then, they remove instances from CFD whose predicted scores from the naive classifier are $\geq$0.99999 or $\leq$0.00001 for the happiness attribute, and $\leq$0.05 for the anger attribute. Finally, they balance the proportions of happiness/anger instances between males and females by removing some images labeled as the majority class (*i.e.*, happy female images for the happiness attribute). We denominate the evaluation subset for the happiness attribute as CFD (Happiness), and for the anger attribute CFD (Anger).

3.2 Proposed Method

Previous work (Cunningham et al., 1995) suggests that there is no single feature or dimension that determines attractiveness, and that attractiveness is the result of combining several features, which individually represent different aspects of a persons’ face. Moreover, this theory indicates that some facial qualities are perceived as universally (physically) attractive. Similarly, facial expressions can be seen as a multi-signal system (Revina and Emmanuel, 2021), and have been shown to posses universal meaning, regardless of culture and gender (Ekman, 1993, 1976). Based on these premises, we propose a method that combines several models trained on two main concepts: one based on different geometrical traits (*objective annotations*) and another based on human judgment (*subjective annotations*).

In this chapter, we detail our proposed method. First, we describe the annotation methodology (Section 3.2.1) for both the attractiveness classification as well as facial expression recognition tasks. Then, we describe the training procedure (Section 3.2.2), which includes the model architectures, optimizer and hyperparameter choices. Finally,

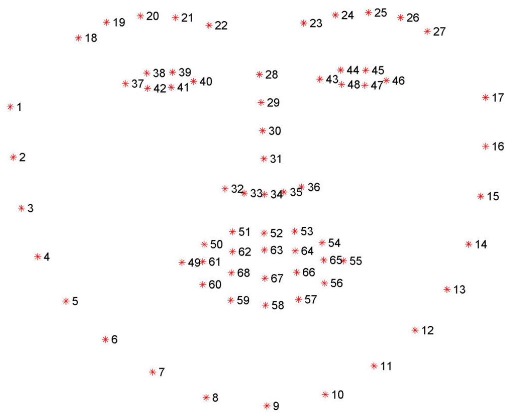


Figure 3.1: Feature points of a facial image from Openface detector (Baltrušaitis et al., 2016).

we introduce the ensemble framework (Section 3.2.3) we use to combine the individual models, each trained in a different definition.

3.2.1 Annotation Methodology

In this section, we describe the annotation methodology we implemented for each task. Since we are proposing the use of *objective* annotations, which are designed individually for each task, each annotation procedure has its peculiarities. Thus, we split the explanation into two parts: first, we detail the annotation procedures of both objective labels as well as the skin color labels for the attractiveness attribute; then, we explain the annotation method for the objective labels of facial expression attribute.

Attractiveness

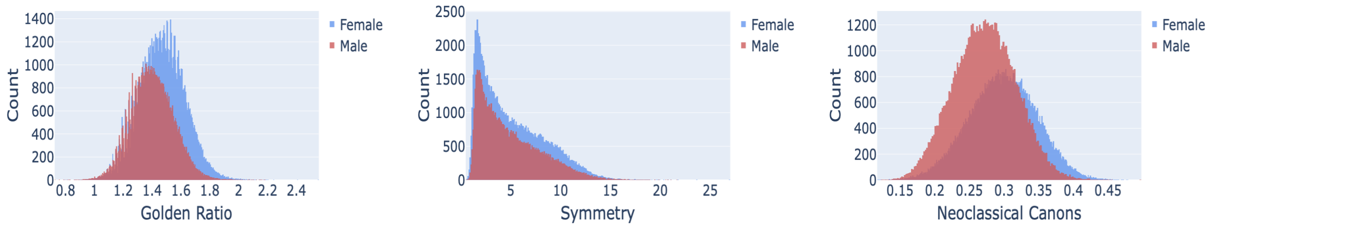
For annotating the objective labels in the attractiveness classification task, we initially extract the facial landmarks using the Openface sofrware (Baltrušaitis et al., 2016), which captures 68 (x, y) coordinates as shown in Figure 3.1. We then map the indexes of the 68 landmarks extracted by Openface to the 29 used in the work of Schmid et al. (2008). During this process, some lateral facial poses (*i.e.*, images in which one side of the face is mostly hidden) are not detected by the face detector. Thus, these images are automatically removed from the training dataset.

Given that the attractiveness measures we are using in this step are based on geometric traits of the face, and to avoid miscalculation, we additionally discard other detected lateral facial poses. Specifically, we consider lateral images, and thus remove them, the ones in which the difference between the distance of the eyes (points 11 and 14 in Figure 2.3) is less than a given threshold. Formally, we test if $d(p_{11}, p_m) - d(p_{14}, p_m) < \beta$, where p_m is the point located in the middle of the face and d is the euclidean distance between the points, as described in Equation 2.6. The ideal value for β is zero as this represents a perfect symmetric face, *i.e.*, the distance from the middle to both sides of the face is equal, thus, the image is not lateral. However, requiring that β be zero is a very hard constraint. We empirically tested several thresholds on thousand images and found that the best one was $\beta = 10$. In other words, for $\beta = 10$ we were able to maintain most of the images of the dataset, while obtaining mostly frontal images. At the end of this process, we kept $137,048$ of the $162,770$ training images (84.19%), and $16,896$ of the $19,867$ validation images (85.04%).

Once we are able to detect all the facial landmarks in frontal images, we calculate the mathematical notion of attractiveness for all dataset images. The mathematical notion may be calculated by the golden ratio, symmetry, or neoclassical canons. It is important to highlight that we do not generate annotations for the sensitive attributes of gender expression and age. In such cases, we simply use the ones provided by the CelebA dataset, *i.e.*, gender was taken from the "Male" property, while age was used from the "Young" property. Thus, age is not a continuous observation, rather, it is defined as a binary attribute. Gender is the gender expression that humans infer when observing the images. Furthermore, the presumed *subjective* annotations are the ones we obtain when using the original attractiveness labels from the CelebA. We solely generate annotations for the *objective* definition of attractiveness based on geometrical traits of the human face, and for the skin color annotations, which we describe later in this section.

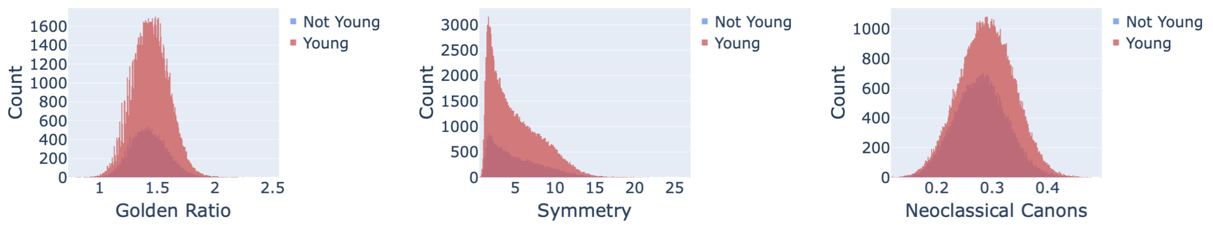
Figure 3.2 shows the dataset distribution for female and male for each calculated metric. Even though the distributions for the same metric seem similar across gender, the high peaks for male and female are slightly different for all of them. Additionally, each metric has its own ideal (target) value. For instance, golden ratio defines the best ratio as 1.618, while symmetry and neoclassical canons define it as 0. We note that neoclassical canons is the curve which possesses the most distant peak from the target value. We hypothesize that this attractiveness metric may be too restrictive, thus not perfectly reflecting the real distribution of the data. This might be related to the fact that neoclassical canons is based on artistic concepts, as mentioned by Farkas et al. (1985).

We depicted the distribution for the age attribute in Figure 3.3. We can visualize that the gap between young ($Y = 1$) and not young ($Y = 0$), for both symmetry and golden ratio, is greater than the one found for the gender attributes. Moreover, we can visualize that in general there exists more images associated with the young sensitive attribute (shown in red) than with the not young (shown in blue). Thus, the distribution across young and not young is less balanced than the distribution of male and female. However, the distribution curves of all objective annotations are still similar across both youth and gender sensitive attributes.



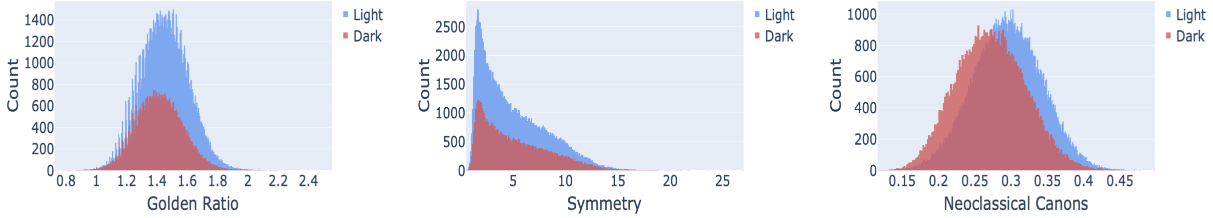
(a) Golden ratio. (b) Symmetry. (c) Neoclassical canons.

Figure 3.2: Histograms showing the data distributions for the attractive attribute stratified by the *gender* attribute.



(a) Golden ratio. (b) Symmetry. (c) Neoclassical canons.

Figure 3.3: Histograms showing the data distributions for the attractive attribute stratified by the *age* attribute.



(a) Golden ratio. (b) Symmetry. (c) Neoclassical canons.

Figure 3.4: Histograms showing the data distributions for the attractive attribute stratified by the *skin color* attribute.

We annotate a third and last sensitive attribute based on skin color. We followed a similar procedure to the work of Sattigeri et al. (2019). CelebA contains multiple images of the same celebrity. Thus, we automatically annotated the average skin color of all the images of each of the celebrities, and propagated the annotation to all of their corresponding images. To reduce the annotation error due to colors not related to skin, *e.g.*, beard or makeup, we selected the pixels from the facial region of the cheeks. Thus, for the left and right sides of the face, we selected the pixels located inside the points 1, 2, 31 and 49, and points 14, 15, 35, 53 from Figure 3.1, respectively. We also added the region inside the points 7, 8, 9, 57, which are close to the chin. We removed points which were considered outliers, *i.e.*, whose colors were had a z-score, defined as $\frac{p-\mu}{\sigma}$, above 3 (standard deviation).

Finally, we calculate the Individual Topology Angle (ITA) on the average value of the selected skin pixels to assign each image a skin tone from the Fitzpatrick skin type scale (Fitzpatrick, 1975). ITA is assessed by colorimetry (Chardon et al., 1991), and it is calculated using the following equation:

$$ITA = \frac{arctan(Lum^{*} - 50)}{b^{*}} \cdot \frac{180}{\pi}, \tag{3.1}$$

where Lum^{*} represents luminance ranging from black (0) to white (100) and b^{*} ranging from yellow to blue. The higher the ITA, the lighter the skin (Osto et al., 2022). ITA skin color types are classified into six groups, from very light to dark skin: very light ($> 55^\circ$), light (41° to $\leq 55^\circ$), intermediate (28° to $\leq 41^\circ$), tan (10° to $\leq 28^\circ$), brown (-30° to $\leq 10^\circ$, and dark ($\leq -30^\circ$). Since we are working with binary attribute, we categorized the types I (very light), II (light), and III (intermediate) as $S = 0$ and the types IV (tan), V (brown), and VI (dark) were categorized as $S = 1$. Figure 3.4 depicts the data distribution for this sensitive attribute. We can visually inspect that for this attribute, the curves of both golden ratio and symmetry distributions are similar to the one obtained for the gender expression attribute. However, we see a predominance of lighter (blue) than darker (red) skin for these two mathematical concepts of beauty. For the neoclassical canons, dark and light skin colors obtain similar distributions, where the peak of lighter skin instances is slightly shifted towards higher values.

Since all of the attractiveness definitions generate a unique continuous value describing the attractiveness of each person, and the CelebA dataset has binary attribute annotations, to obtain consistency we define five different ranges for each mathematical attractiveness measure. These ranges correspond to the amount of variation each metric tolerates. For golden ratio, since its ideal value is 1.618, we define a delta (δ) value that defines the range for which one is considered attractive, *e.g.*, for a hypothetical $\delta = 0.1$, we consider all images that possess golden ratio from 1.518 ($1.618 - \delta$) to 1.718 ($1.618 + \delta$) as attractive. In contrast, since symmetry and neoclassical canons establish the ideal value as 0, and negative values for both metrics are infeasible, we use a threshold t that defines the range of attractive people from 0 to t. For instance, for a symmetry/neoclassical canon of $t = 4$ the person is considered attractive if it contains a proportion between 0 and 4.

Therefore, the higher the δ or the t, the more people will fit into the attractive category ($Y = 1$). Our goal when choosing δ and t is to study the effects of different data distributions in the model's behavior. Thus, at the end, we tested, for each metric (golden ratio, neoclassical canons, and symmetry), five different label distributions, one for each threshold t and δ. Some choices of t and δ are close to the dataset distribution generated by humans, as shown in Tables 4.1, 4.2 and 4.3.

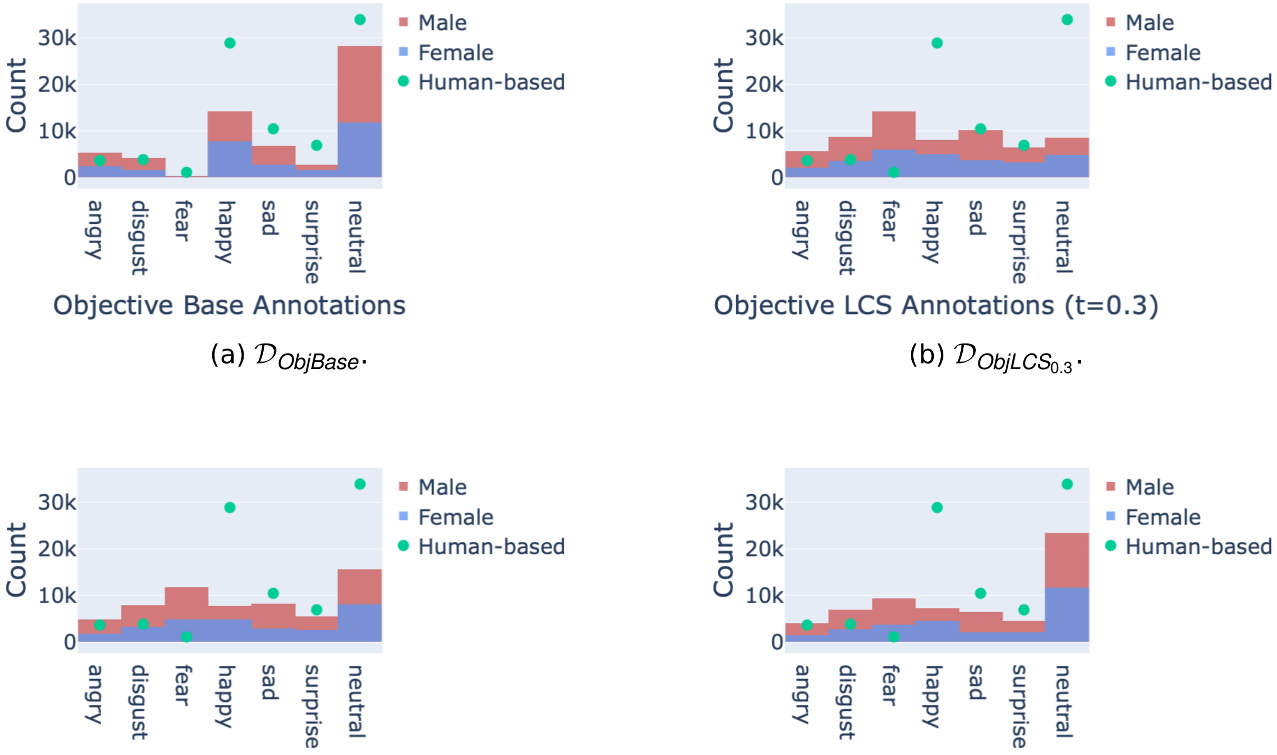
Expression	Action Units
Happiness	6, 12
Sadness	1, 4, 15
Surprise	1, 2, 5, 26
Fear	1, 2, 4, 5, 7, 20, 26
Anger	4, 5, 7, 23
Disgust	9, 15, 17
Neutral	-

Table 3.1: Action Units (AUs) that compose each of the six basic expressions (Ekman, 1993), and the neutral expression which implies no AU is active.

Facial Expression

For the FER system, we use all the pre-processing procedures and datasets provided in the work of Chen and Joo (2021). We highlight the fact that, in their work, the training datasets and the test splits differ. ExpW dataset (Zhang et al., 2015, 2018b) was used for training the classifier for happiness detection, while AffectNet (Mollahosseini et al., 2017) was used for training the classifier for anger detection. Two modified versions with a balanced subset of the CFD dataset (Ma et al., 2015) were used as evaluation sets, each one designed for a specific attribute (happiness/anger). Moreover, in their work, all models are treated as a binary classification, *i.e.*, happy($Y = 1$)/unhappy($Y = 0$) and angry($Y = 1$)/non-angry ($Y = 0$). The negative classes ($Y = 0$) for both happiness and anger classifiers are defined as all the instances that are not annotated as happy or angry, respectively, in the original human-based labels.

We follow a similar annotation procedure for both tasks. For generating the objective annotations, our first step is composed of extracting the AUs for all images in the training dataset. We extract both the presence (as a binary feature) as well as the intensity (as a float feature) of each AU. A complete table of which AUs compose each of the basic expressions is shown in Table 3.1. Next, we create two algorithms, one which we named *ObjBase*, based mainly on the detection of AUs, and a second named *ObjLCS*, which is also based on intensities. For the first one, we simply test whether the combination of AUs for a specific facial expression exists, *i.e.*, whether all the AUs that compose an expression (Ekman and Friesen, 1978) are active for a particular image. For instance, the sadness emotion is composed of the AUs 1, 4 and 15. Therefore, in order for an image to be annotated as ‘sad’, using the ObjBase algorithm, all three AUs need to be active. Thus, we only annotate the image as containing a particular facial expression if all the AUs that compose that expression are active. In case two facial expressions are detected, which happens in 12% of images for the happy attribute and 7% for the anger attribute, we average the intensity of the AUs that compose the facial expressions that were detected. Finally, we annotate the one which contains the highest average since this indicates the AUs that compose this expressions are the predominant ones.



(a) $\mathcal{D}_{ObjBase}$.

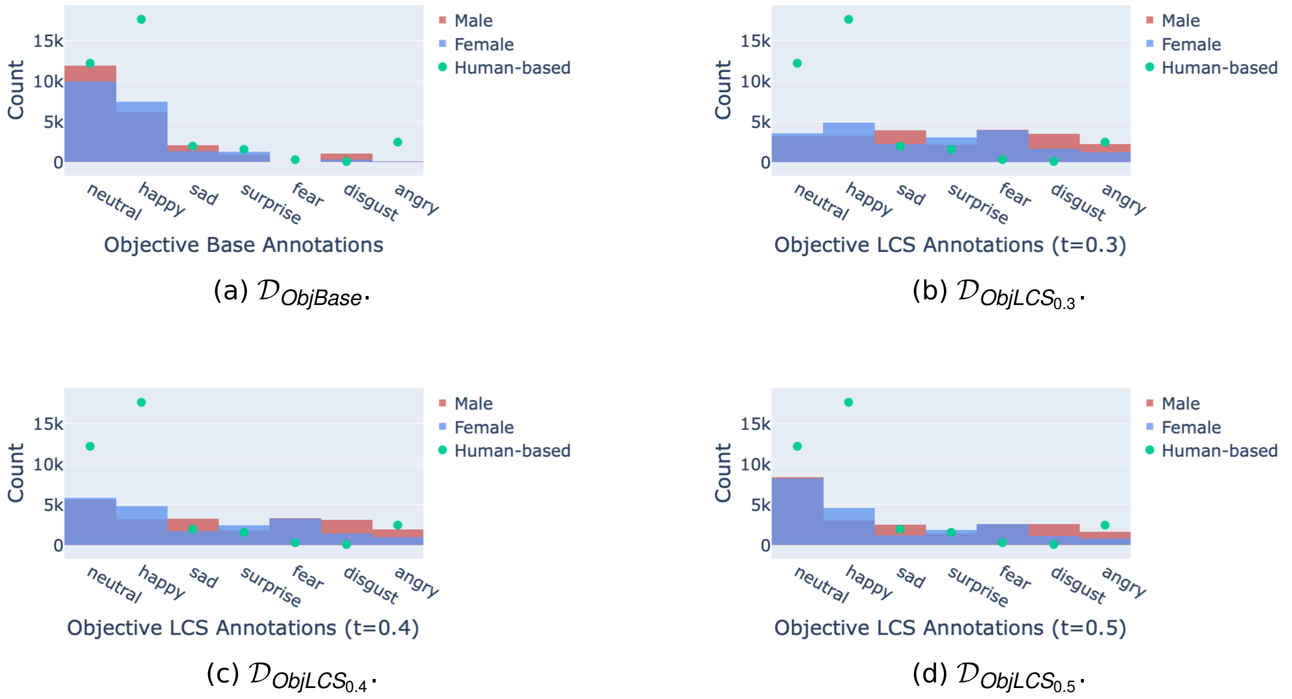
(b) $\mathcal{D}_{ObjLCS_{0.3}}$.

(c) $\mathcal{D}_{ObjLCS_{0.4}}$.

(d) $\mathcal{D}_{ObjLCS_{0.5}}$.

Figure 3.5: Histograms showing the data distributions per gender for the *happiness* attribute of the ExpW dataset (Zhang et al., 2015, 2018b). Green dots represent the frequency of each facial expression for the human-based annotations.

However, this algorithm was designed to only annotate an expression if all the AUs that compose that expression are active in the face. Thus, this algorithm creates the requirement that all AUs be active in order to annotate such an expression. This becomes a very hard constraint as the number of AUs that compose an expression grows, which is the case for some expressions, such as 'fear' which is composed of a combination of seven AUs. This can be easily observed in both Figure 3.5a for the happiness attribute, as well as in Figure 3.6a for anger attribute, since almost no instance (or no instance at all, which is the case for the anger attribute) is labeled as containing the 'fear' expression. Hence, next, we propose a second algorithm for annotating the expressions in an objective way. We use the Longest Common Subsequence (LCS) (Velusamy et al., 2011) method, which is a function whose purpose is to count the number of operations required to transform one string (*i.e.*, detected AUs in an image) into another (*i.e.*, AUs that compose a facial expression). Our goal is to find the expression which possesses the AUs closer to the AUs detected in each image, not necessarily requiring that all AUs pertaining to that expression be active. For the 'sadness' expression, this would not require that all three AUs be active simultaneously for the 'sad' expression to be considered as a possible label. Thus, this



(a) $\mathcal{D}_{ObjBase}$.

(b) $\mathcal{D}_{ObjLCS_{0.3}}$.

(c) $\mathcal{D}_{ObjLCS_{0.4}}$.

(d) $\mathcal{D}_{ObjLCS_{0.5}}$.

Figure 3.6: Histograms showing the data distributions per gender for the *anger* attribute of the AffectNet dataset (Mollahosseini et al., 2017). Green dots represent the frequency of each facial expression for the human-based annotations.

method aims to produce a less strict annotation procedure, and a more diverse set of labels.

Using the LCS method, we compare the AUs detected in the image with the AUs that represent each of the six basic emotions as found by previous literature (Mavadati et al., 2013). However, using this method might also select more than one expression for each image. In this case, we follow the work of Peres and Musse (2021) which calculates the euclidean distance between the intensity of the AUs detected by each facial expression selected and the “ideal” ones defined in the literature by Ekman and Friesen (1978). The expression that results in the lowest Euclidean distance is the annotated one. Nevertheless, extracting the neutral expression individually using the LCS algorithm is not possible since we try to approximate the expression which possesses the AUs closer to the active AUs, while the neutral expression implies absolutely no AU is active. To solve this, we use a threshold t that determines a neutral expression if the average intensity of the AUs of the detected facial expression is less than a minimum t. We define three possible thresholds t (0.3, 0.4 and 0.5).

Figures 3.5b, 3.5c and 3.5d show the distribution of using the ObjLCS$_t$ algorithm for annotating the dataset for the happy attribute. The higher the t, the more concentrated the distribution is to the neutral expression (*i.e.*, less evenly distributed among the six facial expressions). The green dots represent the distribution of the expressions for the human-based annotations (H). We note that all thresholds t provide a more distributed labels per facial expression than the one annotated by humans, *i.e.*, for the human-based labels 70.9% of the instances are considered as happy or neutral. Figures 3.6b, 3.6c and 3.6d show the distribution of using the ObjLCS$_t$ algorithm for annotating the dataset for the anger attribute. Again we see that the distribution that has the most spread distribution (across all seven classes) is the one from the dataset with t = 0.3 ($\mathcal{D}_{ObjLCS_{0.3}}$). Moreover, we also note that all thresholds t provide a more distributed labels per facial expression than the one annotated by humans. Thus, all expressions are more well-represented and balanced in the objective annotations. The main difference we note between the anger and happiness distributions is that the AffectNet dataset possesses a more balanced proportion among both female and male genders for all expressions, while ExpW has more male annotations.

3.2.2 Model Training

Following previous work in both attractiveness classification (Ramaswamy et al., 2021; Sattigeri et al., 2019) and facial expression recognition (Chen and Joo, 2021), we consider only binary classification. In other words, all models are treated as a binary classification, *i.e.*, attractive (Y = 1)/non-attractive(Y = 0), happy(Y = 1)/unhappy(Y = 0), and angry(Y = 1)/non-angry(Y = 0). In our experiments for the attractive attribute, we use ResNet-18 (He et al., 2016) as the base architecture, while for the FER, following previous work (Chen and Joo, 2021), we use ResNet-50 pre-trained on ImageNet (Russakovsky et al., 2015). Both architectures are very popular and well-established algorithms for any visual classification task in the ML field (Khan et al., 2020). The inputs of the ResNet-18 model are 128×128 colored images, and 224×224 for the ResNet-50. All models were trained with the cross entropy loss and Adam (Kingma and Ba, 2015) optimizer.

The learning rate (LR) was set to $1e-3$ for the attractiveness classification, and $1e-4$ for the FER. We use LR scheduler for the former one, which reduces the initial value by 0.1 when the validation loss does not improve for 10 concurrent epochs. For the latter one, we reduce the LR by 0.1 every 6 epochs for the initial LR, and every 4 epochs for the rest. In the work of Chen and Joo (2021), they use two different methodologies for sampling instances: for the happiness attribute, they randomly sample 20,000 instances, while for the anger attribute, they sample the instances so that they are balanced for each gender and AU combination. Since happiness is part of the main paper (and anger is part of the supplementary material), we decided to only follow the methodology used for the happiness attributes in facial expressions, *i.e.*, randomly sample the instances. After the training process, we end up with four categories of models for the attractive attribute, each trained on different ground-truth labels: human-based (1) CelebA annotations, geometrically-based (2) golden ratio, (3) symmetry, and (4) neoclassical canons.

For the FER, we end up with three categories of models: (1) human annotations, (2) AUs base, and (3) AUs with LCS. Thus, models in different categories learn different objective functions.

3.2.3 Ensemble

The last main step in our proposed method is combining models trained on different perspectives. This step has two main motivations: (1) most recent approaches replace several human decision-makers with a single algorithm, such as COMPAS for recidivism risk estimation in the U.S. (Angwin et al., 2016). However, in high-stake real-world applications, the decision is taken from multiple human beings. Thus, we argue that one could introduce diversity into machine decision making by instead training a collection of algorithms, each capturing a different perspective about the problem solution, and then combining their decisions in some ensemble manner (*e.g.*, simple or weighted majority voting); our other motivation is (2) the rich literature on ensemble learning, where a combination of a diverse ensemble of predictors have been shown to outperform single predictors on a variety of tasks (Brown et al., 2005). Moreover, some studies argued and demonstrated that one can obtain a fair model by individually combining them into an ensemble (Grgić-Hlača et al., 2017; Bhaskaruni et al., 2019). Specifically, Grgić-Hlača et al. (2017) theoretically showed that, compared to a single decision-making model, a diverse ensemble can both achieve better fairness in terms of uniformilly distributing resources among users, as well as achieve a better accuracy-fairness trade-off.

In this work, we implement bagging, which often considers (a) homogeneous models, *i.e.*, trained using the same architecture; (b) learns each model independently from each other in parallel; (c) combines them following some kind of deterministic averaging process. Our goal when choosing this type of weighting procedure is to analyze different possible combinations of the individual models, each obtaining a different influence (*i.e.*, weight) on the final ensemble. Thus, after individually training each model, we combine them using the following weighted process for each instance X_i of the test set:

$$f(X_i) = \sum_{m=1}^{M} \alpha_m \cdot o_m(X_i), \tag{3.2}$$

where $f(X_i)$ is the ensemble prediction for the instance i of the input features X, M is the number of individual models we combine, o_m is the output of the m_{th} model for instance X_i, and α_m represents the weight that the m_{th} model has in the final ensemble output of $f(X_i)$. Hence, each model will have a different influence in the final decision.

4. RESULTS

In this section we describe our main results for both attractiveness (Section 4.1) and FER (Section 4.2) tasks. Each follow the same structure of sections: (1) we first report the distribution of the generated datasets, *i.e.*, the database we objectively annotate using the instances provided by CelebA (Liu et al., 2015), ExpW (Zhang et al., 2015, 2018b) and AffectNet (Mollahosseini et al., 2017); (2) we then show the result of the individual models on the test sets; (3) next, we show the results when combining individual models, trained on different subjective and objective notions, into an ensemble; (4) subsequently, we compare our ensemble results with previous literature; (5) finally, we use an explainability method [1] to understand which factors contributed the most for the predictions made by each of the models that compose the final ensemble.

4.1 Attractiveness

In this section we describe the results for the attractiveness classification task.

4.1.1 Dataset Distribution

Our goal when generating different δ and t choices was to study the impact that different data distributions have in the model's behavior. In other words, in this section, we aim to analyze whether slightly altering the dataset distribution heavily affects the models' behavior. We highlight the fact that the higher the δ and t are the more images are considered as attractive. In Table 4.1 we show the distribution of the dataset regarding the new attractiveness measures for each attractiveness range δ or threshold t, also scattered across the sensitive attribute of gender expression, *i.e.* male (S = 1) and female (S = 0). We also added the new distribution of the CelebA dataset (human-based, $\mathcal{D}_H$) when removing lateral facial poses from the training set.

We first observe in Table 4.1 that the target attribute has a distribution close to to the one obtained from the human-labels (53% for Y = 1, 47% for Y = 0) in at least one option of δ and t for all attractiveness metrics. For instance, the range δ = 0.20 for golden ratio has 54% attractive people, the thresholds t = 4.6, for symmetry, t = 0.29 for neoclassical canons, have 54% and 52% attractive people, respectively. We also note that the distribution for the gender attribute varies according to the target attribute and threshold, *i.e.*, when the target attribute is close to the distribution from $\mathcal{D}_H$, the

[1] We used the implementation of saliency method from https://captum.ai

$\mathcal{D}$	δ or t	$Y = 1$	$S = 1$	$S = 0$	$Y = 0$	$S = 1$	$S = 0$
$\mathcal{D}_{GR}$	0.17	46%	32.5%	67.5%	54%	48.9%	51.1%
	0.18	48%	32.9%	67.1%	52%	49.3%	50.7%
	0.19	51%	33.4%	66.6%	49%	49.7%	50.3%
	0.20	54%	33.8%	66.2%	46%	50.2%	49.8%
	0.21	56%	34.2%	65.8%	44%	50.6%	49.4%
$\mathcal{D}_{Sym}$	4.0	47%	41.9%	58.1%	53%	40.9%	59.1%
	4.2	50%	42.0%	58.0%	50%	40.8%	59.2%
	4.4	52%	42.1%	57.9%	48%	40.6%	59.4%
	4.6	54%	42.2%	57.8%	46%	40.4%	59.6%
	4.8	56%	42.3%	57.8%	44%	40.3%	59.7%
$\mathcal{D}_{NC}$	0.26	29%	56.5%	43.5%	71%	35.2%	64.8%
	0.27	36%	54.8%	45.2%	64%	33.7%	66.3%
	0.28	44%	53.2%	46.8%	56%	32.1%	67.9%
	0.29	52%	51.7%	48.3%	48%	30.1%	69.9%
	0.30	60%	49.8%	50.2%	40%	28.2%	71.8%
$\mathcal{D}_H$	-	53%	23.3%	76.7%	47%	61.7%	38.3%

Table 4.1: Dataset distribution for the sensitive attribute of *gender*, *i.e.*, male ($S = 1$) and female ($S = 0$). We display each attractiveness notion per range δ or threshold t for both attractive ($Y = 1$) and not attractive ($Y = 0$). $\mathcal{D}_{GR}$, $\mathcal{D}_{Sym}$, $\mathcal{D}_{NC}$ and $\mathcal{D}_H$ correspond to datasets that capture different attractiveness definitions: golden ratio (GR), symmetry (Sym), neoclassical canons (NC), and human (H) perception, respectively. We highlight that the distribution is similar across train, validation and test sets.

distribution of male/female is also close to 50%. The only exception are the human-based ($\mathcal{D}_H$) labels, which have a distribution of ≈ 20 ($S = 1$) to 50% ($S = 0$) for $Y = 1$ and ≈ 60 ($S = 1$) to 40% ($S = 0$) for $Y = 0$.

In Table 4.2 we present the same dataset distribution, but with respect to the sensitive attribute of age, *i.e.* young ($S = 1$) and not young ($S = 0$). We first note that young/not young have more skewed distributions for both attractive and not attractive options, even for the original CelebA annotations ($\mathcal{D}_H$), than the one we obtained for the gender expression. In other words, we can visualize that the proportion of young ($S = 1$) and not young ($S = 0$) for both attractive ($Y = 1$) and not attractive ($Y = 0$) for all the objective annotations, regardless of δ or t, is around $75 - 80\%$ to $20 - 25\%$, respectively. However, the discrepancy can be even higher for the human-based ($\mathcal{D}_H$) annotation, which has 92.9% of the attractive people labeled as young. We can also conclude that the gap in distribution of young and not young for the different δ and t definitions of the same objective definition is smaller, *e.g.*, for $\mathcal{D}_{GR}$ from $\delta = 0.17$ to $\delta = 0.21$ the percentage of attractive young people only varies from 79.2% yo 78.7%.

Finally, in Table 4.3 we show the dataset distribution with respect to the different skin color annotations we generate, *i.e.*, dark ($S = 1$) and light ($S = 0$) skin. For the attractive labels ($Y = 1$), all the objective annotations for light/dark skin color are more evenly distributed than the ones obtained from the human-based ($\mathcal{D}_H$) labels. This does

$\mathcal{D}$	δ or t	$Y = 1$	$S = 1$	$S = 0$	$Y = 0$	$S = 1$	$S = 0$
$\mathcal{D}_{GR}$	0.17	46%	79.2%	20.8%	54%	75.0%	25.0%
	0.18	48%	79.1%	20.9%	52%	74.9%	25.1%
	0.19	51%	78.9%	21.1%	49%	74.8%	25.2%
	0.20	54%	78.9%	21.1%	46%	74.7%	25.3%
	0.21	56%	78.7%	21.3%	44%	74.6%	25.4%
$\mathcal{D}_{Sym}$	4.0	47%	77.4%	22.6%	53%	76.5%	23.5%
	4.2	50%	77.3%	22.7%	50%	76.5%	25.5%
	4.4	52%	77.3%	22.7%	48%	76.6%	23.4%
	4.6	54%	77.2%	22.8%	46%	76.5%	23.5%
	4.8	56%	77.2%	22.7%	44%	76.6%	23.4%
$\mathcal{D}_{NC}$	0.26	29%	72.8%	27.2%	71%	78.6%	21.4%
	0.27	36%	73.1%	26.9%	64%	79.1%	20.9%
	0.28	44%	73.6%	26.4%	56%	79.6%	20.4%
	0.29	52%	74.8%	26.2%	48%	80.3%	19.7%
	0.30	60%	74.1%	25.9%	40%	81.1%	41.1%
$\mathcal{D}_H$	-	53%	92.9%	7.1%	47%	58.9%	41.1%

Table 4.2: Dataset distribution for the sensitive attribute of *age*, *i.e.*, young (S = 1) and not young (S = 0). We display each attractiveness notion per range δ or threshold t for both attractive (Y = 1) and not attractive (Y = 0). $\mathcal{D}_{GR}$, $\mathcal{D}_{Sym}$, $\mathcal{D}_{NC}$ and $\mathcal{D}_H$ correspond to datasets that capture different attractiveness definitions: golden ratio (GR), symmetry (Sym), neoclassical canons (NC), and human (H) perception, respectively. We highlight that the distribution is similar across train, validation and test sets.

not happen in the same proportion for the non-attractive (Y = 0) annotations, where the original labels have a better distribution across this sensitive attribute. Overall, from all the sensitive attributes we consider in this part of the work (gender expression, age, and skin color), we obtained a more balanced arrangement of attractive and non-attractive labels for the objective annotations than the human-based annotations.

4.1.2 Individual Models

In this section, we describe the results when individually training the models. Specifically, we train one model for each δ and t of each attractiveness definition. Thus, in total, we obtained 15 models trained on objective definitions of attractiveness. Our goal is to verify whether testing models trained on objective perspectives on the CelebA test set actually reduces the fairness metric compared to the model originally trained on subjective annotations. We show the results when evaluating all models on the human-based attractiveness concept, using CelebA original test set annotations. We also show the baseline result, which is trained on the original CelebA training labels ($\mathcal{D}_H$). Each table shows an average result across three different runs, each with a different and randomly picked seed (3, 18, and 54). We highlight the fact that all the models were trained only

$\mathcal{D}$	δ or t	$Y = 1$	$S = 1$	$S = 0$	$Y = 0$	$S = 1$	$S = 0$
$\mathcal{D}_{GR}$	0.17	46%	26.7%	73.3%	54%	33.7%	66.3%
	0.18	48%	26.9%	73.1%	52%	33.8%	66.2%
	0.19	51%	27.1%	72.9%	49%	34.0%	66.0%
	0.20	54%	27.2%	72.8%	46%	34.2%	65.8%
	0.21	56%	27.4%	72.6%	44%	34.4%	65.6%
$\mathcal{D}_{Sym}$	4.0	47%	29.4%	70.6%	53%	31.4%	68.6%
	4.2	50%	29.5%	70.5%	50%	31.4%	68.6%
	4.4	52%	29.6%	70.4%	48%	31.4%	68.6%
	4.6	54%	29.6%	70.4%	46%	31.5%	68.5%
	4.8	56%	29.6%	70.4%	44%	31.5%	68.5%
$\mathcal{D}_{NC}$	0.26	29%	43.9%	56.1%	71%	25.0%	75.0%
	0.27	36%	41.9%	58.1%	64%	24.0%	76.0%
	0.28	44%	40.0%	60.0%	56%	23.0%	77.0%
	0.29	52%	38.3%	61.7%	48%	22.0%	78.0%
	0.30	60%	36.8%	63.2%	40%	21.0%	79.0%
$\mathcal{D}_{H}$	-	53%	18.0%	82.0%	47%	44.6%	55.4%

Table 4.3: Dataset distribution for the sensitive attribute of *skin color*, *i.e.*, dark ($S = 1$) and light ($S = 0$). We display each attractiveness notion per range δ or threshold t for both attractive ($Y = 1$) and not attractive ($Y = 0$). $\mathcal{D}_{GR}$, $\mathcal{D}_{Sym}$, $\mathcal{D}_{NC}$ and $\mathcal{D}_{H}$ correspond to datasets that capture different attractiveness definitions: golden ratio (GR), symmetry (Sym), neoclassical canons (NC), and human (H) perception, respectively. We highlight that the distribution is similar across train, validation and test sets.

using the frontal images, *i.e.*, all models were trained on the same set of images, however, each one used a different annotation procedure.

Table 4.4 shows the results for models trained on golden ratio, symmetry, and neocanons, respectively, evaluated on CelebA test set for the sensitive attribute *male*. We first note a trade-off between accuracy (‘Overall’ Accuracy column) and fairness (ΔEoO column), as previously discussed in the literature (Haas, 2019). This is especially the case for models trained on objective annotations (*i.e.*, $\mathcal{D}_{GR}$, $\mathcal{D}_{Sym}$ and $\mathcal{D}_{NC}$), compared to the one trained on CelebA ($\mathcal{D}_{H}$), which obtained a better fairness result (lower ΔEoO) but close to random overall accuracy (50% in a binary classification problem). However, the low accuracy is expected since they were not trained to capture subjective human-like patterns, instead, they were trained to detect mathematical definitions of attractiveness. Furthermore, we notice that the ΔEoO is much lower on the models based on geometrical traits than the one trained on human-based labels. In other words, *all* the models trained on geometrical concepts of attractiveness, for the gender expression, are much less discriminatory in the chosen fairness metric than the one trained on subjective notions, regardless of the choice of δ or t. Lastly, we note that for the same attractiveness definition all models obtain similar accuracy and fairness values, *i.e.*, the choice of δ or t does not heavily impact the individual results.

		Accuracy					
$\mathcal{D}$	δ or t	Overall	$S = 1$	$S = 0$	ΔTPR	ΔFPR	ΔEoO
$\mathcal{D}_{GR}$	0.17	0.555	0.517	0.614	0.162	0.221	0.162
	0.18	0.559	0.533	0.601	0.144	0.218	**0.144**
	0.19	0.557	0.533	0.595	0.160	0.231	0.160
	0.20	0.558	0.543	0.582	0.148	0.219	0.148
	0.21	0.553	0.551	0.555	0.156	0.216	0.156
$\mathcal{D}_{Sym}$	4.0	0.518	0.482	0.574	0.083	0.008	**0.083**
	4.2	0.518	0.488	0.565	0.085	0.013	0.085
	4.4	0.516	0.491	0.557	0.091	0.017	0.091
	4.6	0.514	0.495	0.545	0.090	0.019	0.090
	4.8	0.515	0.502	0.536	0.093	0.015	0.093
$\mathcal{D}_{NC}$	0.26	0.443	0.390	0.529	0.132	0.150	**0.132**
	0.27	0.436	0.420	0.462	0.143	0.164	0.143
	0.28	0.428	0.429	0.427	0.168	0.190	0.168
	0.29	0.432	0.465	0.381	0.158	0.179	0.158
	0.30	0.439	0.505	0.333	0.165	0.190	0.165
$\mathcal{D}_H$	-	0.807	0.796	0.825	0.193	0.275	0.193

Table 4.4: Results of individual models for the sensitive attribute *gender*, *i.e.*, $S = 1$ represents male, while $S = 0$ represents female. Each model was trained on different attractiveness notions and evaluated on CelebA original test set. We show the average results across three different seeds. $\mathcal{D}_{GR}$, $\mathcal{D}_{Sym}$, $\mathcal{D}_{NC}$ and $\mathcal{D}_H$ correspond to the different datasets that the models were trained on, each based on a different attractiveness notion: golden ratio, symmetry, neoclassical canons, and human perception, respectively. We highlighted the best and underlined the worst average ΔEoO results.

Likewise, Table 4.5 depicts the results for the sensitive attribute *young*. As it is observed in the gender expression, there is a trade-off between the overall accuracy and the fairness metric for all models regarding the age attribute. Moreover, for the fairness metric we visualize an even bigger gap between the model trained on human-based labels, which obtained $\Delta EoO = 0.189$, and the models trained on objective definitions of attractiveness, which obtained a maximum ΔEoO of 0.099 for the model trained on *GR* with $\delta = 0.17$. Thus, again we observe that *all* the models trained with objective definitions of attractiveness obtain lower discriminatory behavior, measured by ΔEoO, than the model trained on subjective notions. Finally, Table 4.6 shows the results for the *skin color* attribute. The results follow a similar pattern, where we also observe the trade-off between utility (overall accuracy) and discrimination (fairness metric). However, we note that the results for the neoclassical canon annotations provided a ΔEoO slightly higher than the baseline, *i.e.*, model trained with human annotations.

Therefore, from the results shown above, we can conclude that slightly altering δ and t does not have a huge impact on the models' outcome, for both accuracy and ΔEoO. Moreover, in general, individually training models on the geometric notions of attractiveness improve the fairness metrics when tested on subjective annotations, *i.e.*, CelebA test set. This supports our claim that models trained to perceive mathematical no-

		Accuracy					
$\mathcal{D}$	δ or t	Overall	$S = 1$	$S = 0$	ΔTPR	ΔFPR	ΔEoO
$\mathcal{D}_{GR}$	0.17	0.555	0.627	0.532	0.099	0.056	0.099
	0.18	0.559	0.604	0.545	0.092	0.059	0.092
	0.19	0.557	0.597	0.544	0.096	0.061	0.096
	0.20	0.558	0.577	0.552	0.089	0.055	0.089
	0.21	0.553	0.542	0.556	0.086	0.046	**0.086**
$\mathcal{D}_{Sym}$	4.0	0.518	0.572	0.500	0.049	0.018	0.049
	4.2	0.518	0.563	0.503	0.045	0.017	0.045
	4.4	0.516	0.551	0.505	0.041	0.018	**0.041**
	4.6	0.514	0.534	0.508	0.043	0.025	0.043
	4.8	0.515	0.520	0.514	0.048	0.024	0.048
$\mathcal{D}_{NC}$	0.26	0.443	0.600	0.393	0.015	0.021	**0.015**
	0.27	0.436	0.516	0.411	0.030	0.019	0.030
	0.28	0.428	0.468	0.415	0.030	0.002	0.030
	0.29	0.432	0.394	0.444	0.047	0.019	0.047
	0.30	0.439	0.322	0.476	0.054	0.039	0.054
$\mathcal{D}_{H}$	-	0.807	0.863	0.789	0.189	0.254	0.189

Table 4.5: Results of individual models for the sensitive attribute *age*, *i.e.*, $S = 1$ represents young, while $S = 0$ represents people of age. Each model was trained on different attractiveness notions and evaluated on CelebA original test set. We show the average results across three different seeds. $\mathcal{D}_{GR}$, $\mathcal{D}_{Sym}$, $\mathcal{D}_{NC}$ and $\mathcal{D}_{H}$ correspond to the different datasets that the models were trained on, each based on a different attractiveness notion: golden ratio, symmetry, neoclassical canons, and human perception, respectively. We highlighted the best and underlined the worst average ΔEoO results.

tions of attractiveness in fact obtain a lower discriminatory behavior than the ones trained on subjective notions. We showed this for a variety of three different sensitive attributes. However, even though our goal is to add the fairness constraint to the unfair decision-making process, we do not wish to reduce the accuracy to a random-choice level since this results in a useless model that would be misclassifying half the instances. Thus, we next combine all models into an ensemble. Our intuition is that these models, once combined, will produce a final ensemble with high accuracy and low fairness measure.

4.1.3 Ensemble Model

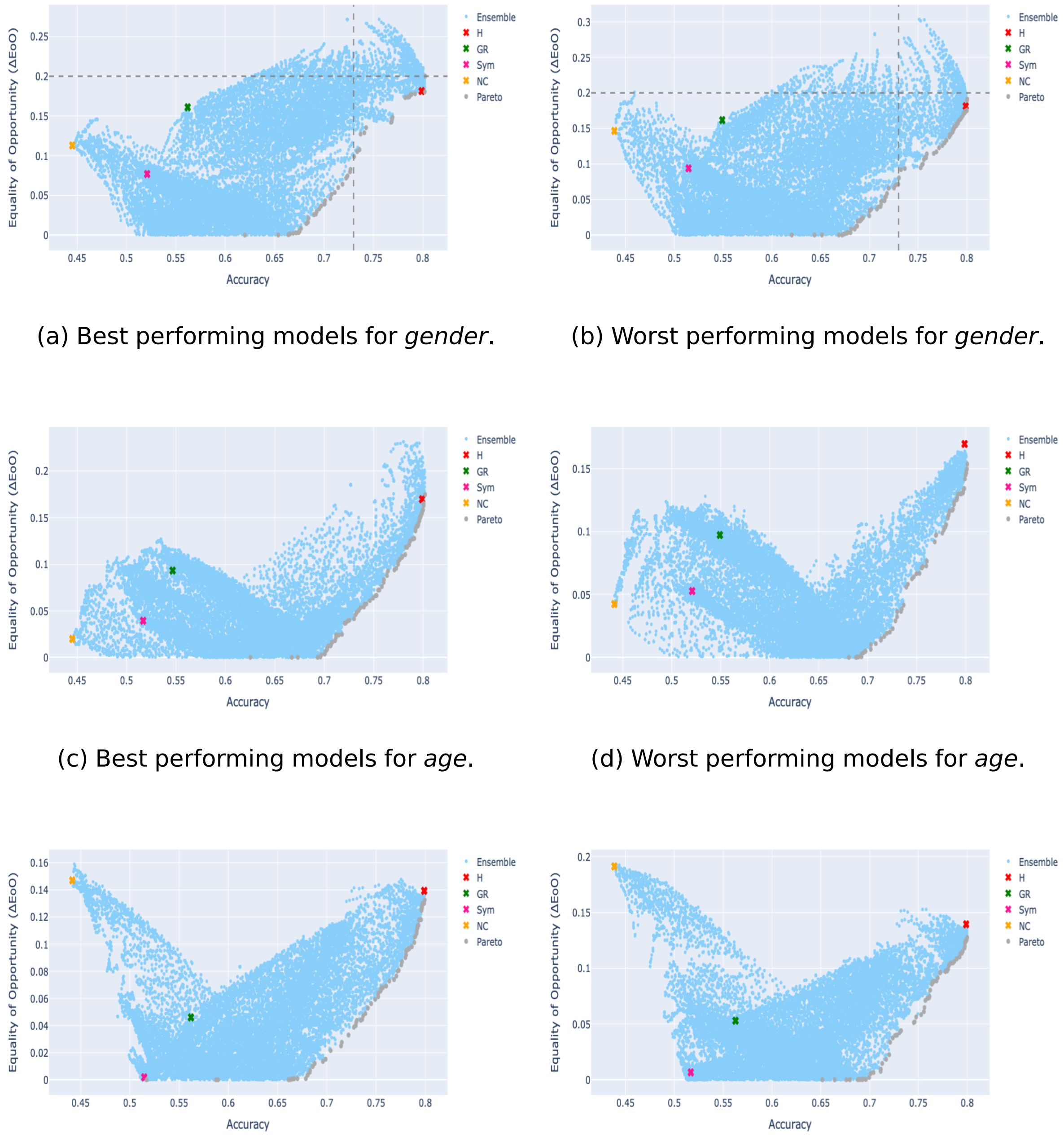
In this section we analyze the impact of combining simple learners into a single complex and diverse model. The ensemble in this section combines four models, each previously trained on a different notion of attractiveness, *i.e.*, (1) Golden Ratio (*GR*), (2) Symmetry (*Sym*), (3) Neoclassical Canons (*NC*), and (4) CelebA (human-based labels, *H*). Since we have several t and δ choices per model, we chose the ones which performed best and worst with respect to the fairness metric (ΔEoO) on CelebA test set (Tables 4.1, 4.2 and 4.3). For the sensitive attribute *male*, the best and worst performing models were

		Accuracy					
$\mathcal{D}$	δ or t	Overall	$S = 1$	$S = 0$	ΔTPR	ΔFPR	ΔEoO
$\mathcal{D}_{GR}$	0.17	0.555	0.534	0.597	0.042	0.081	0.042
	0.18	0.559	0.548	0.584	0.038	0.083	**0.038**
	0.19	0.557	0.548	0.576	0.047	0.085	0.047
	0.20	0.558	0.554	0.566	0.044	0.085	0.044
	0.21	0.553	0.556	0.546	0.040	0.089	0.040
$\mathcal{D}_{Sym}$	4.0	0.518	0.494	0.567	0.008	0.031	0.007
	4.2	0.518	0.497	0.561	0.005	0.029	0.006
	4.4	0.516	0.499	0.552	0.008	0.028	0.008
	4.6	0.514	0.501	0.543	0.006	0.031	0.006
	4.8	0.515	0.507	0.533	0.005	0.030	**0.005**
$\mathcal{D}_{NC}$	0.26	0.443	0.417	0.499	0.173	0.213	0.173
	0.27	0.436	0.436	0.438	0.183	0.218	0.183
	0.28	0.428	0.437	0.409	0.201	0.224	0.201
	0.29	0.432	0.463	0.368	0.191	0.209	0.191
	0.30	0.439	0.490	0.333	0.160	0.173	**0.160**
$\mathcal{D}_{H}$	-	0.807	0.804	0.813	0.154	0.179	0.154

Table 4.6: Results of individual models for the sensitive attribute *skin color*, *i.e.*, $S = 1$ represents dark skin, while $S = 0$ represents people of light skin. Each model was trained on different attractiveness notions and evaluated on CelebA original test set. We show the average results across three different seeds. $\mathcal{D}_{GR}$, $\mathcal{D}_{Sym}$, $\mathcal{D}_{NC}$ and $\mathcal{D}_{H}$ correspond to the different datasets that the models were trained on, each based on a different attractiveness notion: golden ratio, symmetry, neoclassical canons, and human perception, respectively. We highlighted the best and underlined the worst average ΔEoO results.

$GR_{\delta=0.18}$, $Sym_{t=4}$, $NC_{t=0.26}$, and $GR_{\delta=0.17}$, $Sym_{t=4.8}$, $NC_{t=0.28}$, respectively. In contrast, for the attribute *young*, the best and worst performing models were $GR_{\delta=0.21}$, $Sym_{t=4.4}$, $NC_{t=0.26}$, and $GR_{\delta=0.17}$, $Sym_{t=4}$, $NC_{t=0.3}$, respectively. Lastly, for the *skin color* attribute, the best and worst performing models were $GR_{\delta=0.18}$, $Sym_{t=4.8}$, $NC_{t=0.3}$, and $GR_{\delta=0.19}$, $S_{t=4.4}$, $NC_{t=0.28}$, respectively.

As previously described in Section 3.2.3, we used a weighted combination of the models, *i.e.*, each individual model possesses a different influence over the final prediction. Moreover, the previous results showed an average result across three different runs, each with a different seed (3, 18, and 54). However, when combining the models to form the ensemble we randomly chose one seed (18) for all trained models. Figure 4.1 shows the result of several weighting values per model for all the sensitive attribute. We show the result of each ensemble with respect to accuracy (x axis) and ΔEoO (y axis). Each blue dot illustrates the result of one ensemble model (one weighted combination) of the four models. We varied the weight of each individual (base) model from 0 to 1 with steps 0.1. Thus, at the end, we obtain more than $10,000$ possible combinations. We removed the combination of all the weights 0, *i.e.*, all the models having 0 influence, as this implies having no model at all.



(a) Best performing models for *gender*.

(b) Worst performing models for *gender*.

(c) Best performing models for *age*.

(d) Worst performing models for *age*.

(e) Best performing models for *skin color*.

(f) Worst performing models for *skin color*.

Figure 4.1: Results of different weighting values to compose the final ensemble for the attractive attribute. Plots on the first and second column correspond to the best and worst individual models with respect to ΔEoO, respectively. We show the result of each ensemble with respect to accuracy (x axis) and ΔEoO (y axis) for the sensitive attributes *gender*, *age*, and *skin color*. Light blue dots represent the ensemble models, *i.e.*, combining different definitions of attractiveness; red, green, pink and orange crosses indicate the model trained with only CelebA (H), golden ratio (GR), symmetry (Sym) and neoclassical canons (NC) annotations, respectively; finally, the gray dots represent the Pareto analysis.

To best understand the results over the baseline, we also plotted the result of models trained with a single attractiveness definition. Thus, the model trained only on human-based CelebA annotations (*H*) is shown in red, and the ones trained only with mathematical concepts of attractiveness, such as golden ratio (*GR*), symmetry (*Sym*) and neoclassical canons (*NC*), are shown in green, pink, and orange, respectively. The gray dots represent the Pareto analysis, which is based on Pareto efficiency (Iancu and Trichakis, 2014). Pareto-optimal solution in multi-objective optimization delivers optimized performance across different objectives (Iancu and Trichakis, 2014). In this work, we wish to optimize for both accuracy and fairness. Thus, the optimal solutions when maximizing accuracy and minimizing ΔEoO are the ones shown in gray. In the case of the sensitive attribute *male* (Figure 4.1a), we also added the result of the previous state-of-the-art model in gray dotted vertical and horizontal lines. It obtains ΔEoO = 0.2 and overall accuracy of $\approx$ 0.73. Since we do not have a baseline result for the other sensitive attributes, we remove these lines for Figures 4.1c, 4.1d, 4.1e and 4.1f.

We first observe that, with the exception of the sensitive attribute *skin color*, all plots have the model trained on human-based annotations (red cross) as the worst individual result (of all crosses) regarding ΔEoO. Nevertheless, for the skin color, the model trained only with neoclassical canons has a similar result as the human-based labels. For all sensitive attribute, both the curves for the best and the worst performing models show comparable results. This suggests that individually combining the best and worst models into an ensemble have approximately the same result regarding overall accuracy and ΔEoO. Moreover, it reinforces our previous finding, in which the choice of δ and t does not have a huge impact on the final result. We can also visualize, through the Pareto analysis of all plots (gray dots in Figure 4.1) that there seems to be a linear relationship between accuracy and ΔEoO. This is specially the case for the range where the overall accuracy is $\approx$ 0.65 until the maximum accuracy reached by the models ($\approx$ 0.8). More specifically, all plots seem to imply that the higher the accuracy, the higher the fairness metric (higher the ΔEoO), which aligns with previous findings in the literature (Corbett-Davies et al., 2017; Kleinberg et al., 2016).

Moreover, from Figure 4.1, we can conclude that the final result is heavily dependent on the weights each base model has in the final ensemble. This can be directly inferred from how scattered the ensemble models are (light blue dots). For instance, from the horizontal gray line, fixed at ΔEoO = 0.2 (Ramaswamy et al., 2021) in Figure 4.1a and Figure 4.1b, it is possible to obtain an overall accuracy from $\approx$ 0.60 to more than $\approx$ 0.8. Simultaneously, it is also possible to obtain an ensemble whose accuracy is 0.73 (Sattigeri et al., 2019), represented by the vertical gray line in both plots, whose ΔEoO varies from $\approx$ 0.1 to less than $\approx$ 0.2. Nonetheless, for all sensitive attributes, we are able to obtain a fairly high accuracy with a lower ΔEoO. The decision upon which ensemble to choose from will depend deeply on the downstream application.

4.1.4 Comparison with Prior Work

Finally, in this section, we compare our method with previous approaches. The goal of this section is to analyze whether our method obtains a better trade-off between accuracy and fairness compared to previous related work. We are unable to provide a comparative study with any other method for the sensitive attributes of age and skin color since previous work lack public results on age, and skin color was annotated by us. Moreover, previous approaches (Sattigeri et al., 2019; Ramaswamy et al., 2021) focused on creating new instances using generative models, and the cost for reproducing their results, in some cases even without the original source code (Sattigeri et al., 2019), is prohibitive. Thus, we focus on comparing our approach for the gender expression.

In order to choose some ensembles over all possible combinations, we opted for considering only the subset of the models at both Pareto boundaries for the curves presented in Figures 4.1a and 4.1b, *i.e.*, we selected models that were present in the Pareto boundaries for *both* plots (best and worse performing models on the *gender expression* attribute). We selected and sorted the ensemble models according to the fairness score ΔEoO, *i.e.*, the lower the better. In Table 4.7 we show the results and compare our method with previous debiasing approaches for the attractive attribute (Sattigeri et al., 2019; Ramaswamy et al., 2021). We selected four models ($Ours_1$, $Ours_2$, $Ours_9$, $Ours_{10}$), two of them performed the best regarding ΔEoO ($Ours_1$, $Ours_2$), and two of them which achieved the best overall accuracy in the top-10 best ensemble models ($Ours_9$, $Ours_{10}$). For completeness, we show all the 10 best performing ensemble models with respect to ΔEoO in Table 4.8.

We first note that the top-2 performing models regarding ΔEoO obtained a metric that is four times lower than the previous state-of-the-art (ΔEoO = 0.05 against ΔEoO = 0.20 from Ramaswamy et al. (2021)), while holding a competitive overall accuracy. Still, the models that obtained the best overall accuracy in the top-10 performing ensembles also have a lower ΔEoO than previous work. Moreover, we observe that the models that have the greater impact on the final ensemble (*i.e.*, higher weights) are the one trained on human-based labels (H) and the one trained on neoclassical canons concept (NC).

We also notice from Table 4.8 that the model trained on GR labels, in general, has the lowest impact on the final ensembles. This could be due to the fact that the model trained on GR annotations results in the second highest fairness metric, just slightly below the one that uses human annotations, while achieving close-to-random accuracy. Since we are evaluating our models on subjective labels (*i.e.*, generated by humans), the models on the Pareto boundaries possibly have a greater weight on the model trained on human labels, which positively contributes to accuracy, than one that does not have this behavior, *i.e.*, does not contribute to accuracy and does not substantially improve fairness.

	Accuracy			ΔEoO
	$S = 0$	$S = 1$	Overall	
Fairness GAN (Sattigeri et al., 2019)	0.71	0.76	0.73	0.23
LSD (Ramaswamy et al., 2021)	0.79	**0.82**	-	0.20
$Ours_1$ ($H = 1.0; GR = 0.0; Sym = 0.4; NC = 0.7$)	0.70	0.73	0.71	**0.05**
$Ours_2$ ($H = 0.8; GR = 0.1; Sym = 0.4; NC = 0.5$)	0.70	0.75	0.72	**0.07**
$Ours_9$ ($H = 0.5; GR = 0.0; Sym = 0.1; NC = 0.3$)	0.77	0.78	**0.78**	**0.14**
$Ours_{10}$ ($H = 0.9; GR = 0.0; Sym = 0.0; NC = 0.1$)	**0.80**	0.80	**0.80**	**0.17**

Table 4.7: Comparison of our method with previous debiasing methods for the attractiveness attribute with respect to the sensitive attribute *gender*, *i.e.*, $S = 1$ represents male, and $S = 0$ represents female. We show the results for the Fairness GAN (Sattigeri et al., 2019), Latent Space De-biasing (LSD) (Ramaswamy et al., 2021), and a set of the ensemble models we obtained from combining individual models trained on different attractiveness definitions, as shown in Figures 4.1a and 4.1b. We also added the weights that each model has on the final ensemble, *e.g.*, model trained on subjective human-based (H) labels, and models trained on objective (GR, Sym, NC) labels. We highlighted the best average results.

	Weights				Accuracy			ΔEoO
	H	GR	Sym	NC	$S = 0$	$S = 1$	Overall	
1	1.0	0.0	0.4	0.7	0.70	0.73	0.71	0.05
2	0.8	0.1	0.4	0.5	0.70	0.75	0.72	0.07
3	1.0	0.1	0.5	0.6	0.71	0.75	0.72	0.07
4	1.0	0.1	0.4	0.7	0.71	0.75	0.73	0.08
5	0.9	0.2	0.5	0.5	0.71	0.76	0.73	0.09
6	1.0	0.2	0.5	0.6	0.71	0.76	0.73	0.09
7	0.7	0.0	0.3	0.4	0.74	0.77	0.75	0.11
8	0.5	0.0	0.2	0.3	0.75	0.77	0.75	0.11
9	0.5	0.0	0.1	0.3	0.77	0.78	0.78	0.14
10	0.9	0.0	0.0	0.1	0.80	0.80	0.80	0.17

Table 4.8: Weights for the models that compose the top-10 performing ensembles for the gender attribute (model trained on subjective human-based (H) labels, and models trained on objective (GR, Sym, NC) labels), as well as their associated metrics, sorted by ΔEoO. We show overall error rate and ΔEoO, as well as the accuracy rate across both male ($S = 1$) and female ($S = 0$). The individual models that compose these final ensembles are the ones that appear in the Pareto boundaries of both Figures 4.1a and 4.1b.

All of our approaches have the lowest ΔEoO, while maintaining significant accuracy compared to previous work. Moreover, we show that all of our metrics are comparable or better than Sattigeri et al. (2019) and Ramaswamy et al. (2021) approaches, both of which incorporate two to three times more synthetic images to the original training dataset. Thus, we obtained a better trade-off between a given fairness metric (ΔEoO) and accuracy compared to other pre-processing approaches for the attractive attribute.

4.1.5 Model Interpretation

In this section, we describe the results when applying an explainability method to all four models that compose the final ensemble. For this analysis, we use the saliency method described in Section 4.1.5, which generates heatmaps that indicate the regions of the image that the model attended to the most to make a specific prediction. In other words, the darker the region, the more the model relied on that region for making a decision. We first begin by computing the average region that each model attended for all the instances in the dataset. We randomly sampled one image from the test set to plot the average saliency.

Results can be visualized in Figure 4.2. We note the difference in the average region attended by each model: while the model trained on subjective human-based annotations uses a bigger and more spread region across the face, which even contains the forehead and cheeks, the ones trained on objective annotations contain a smaller and more concentrated region. Specifically, models trained on objective annotations concentrate their attention mostly in the nose region, which supports the claim that these models are trained to perceive and classify images according to mathematical ratios.

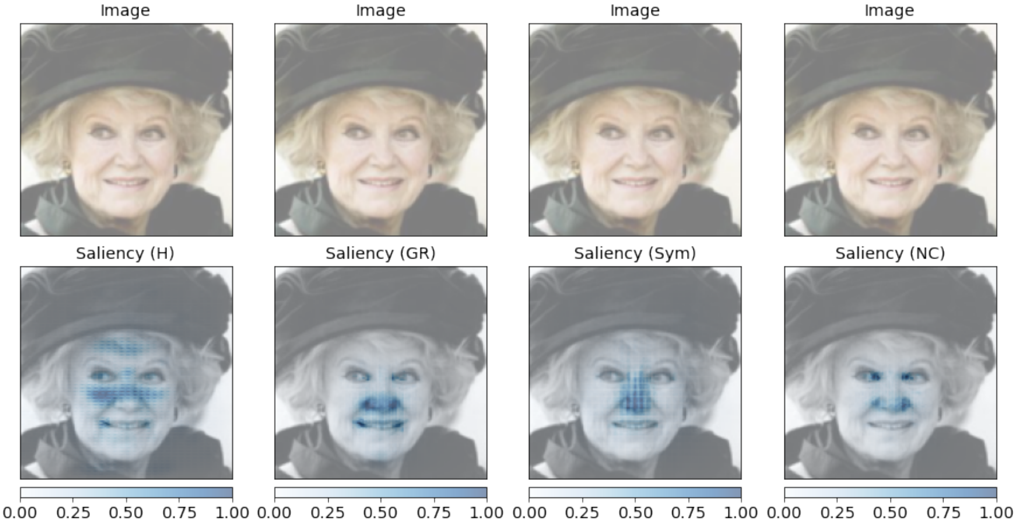
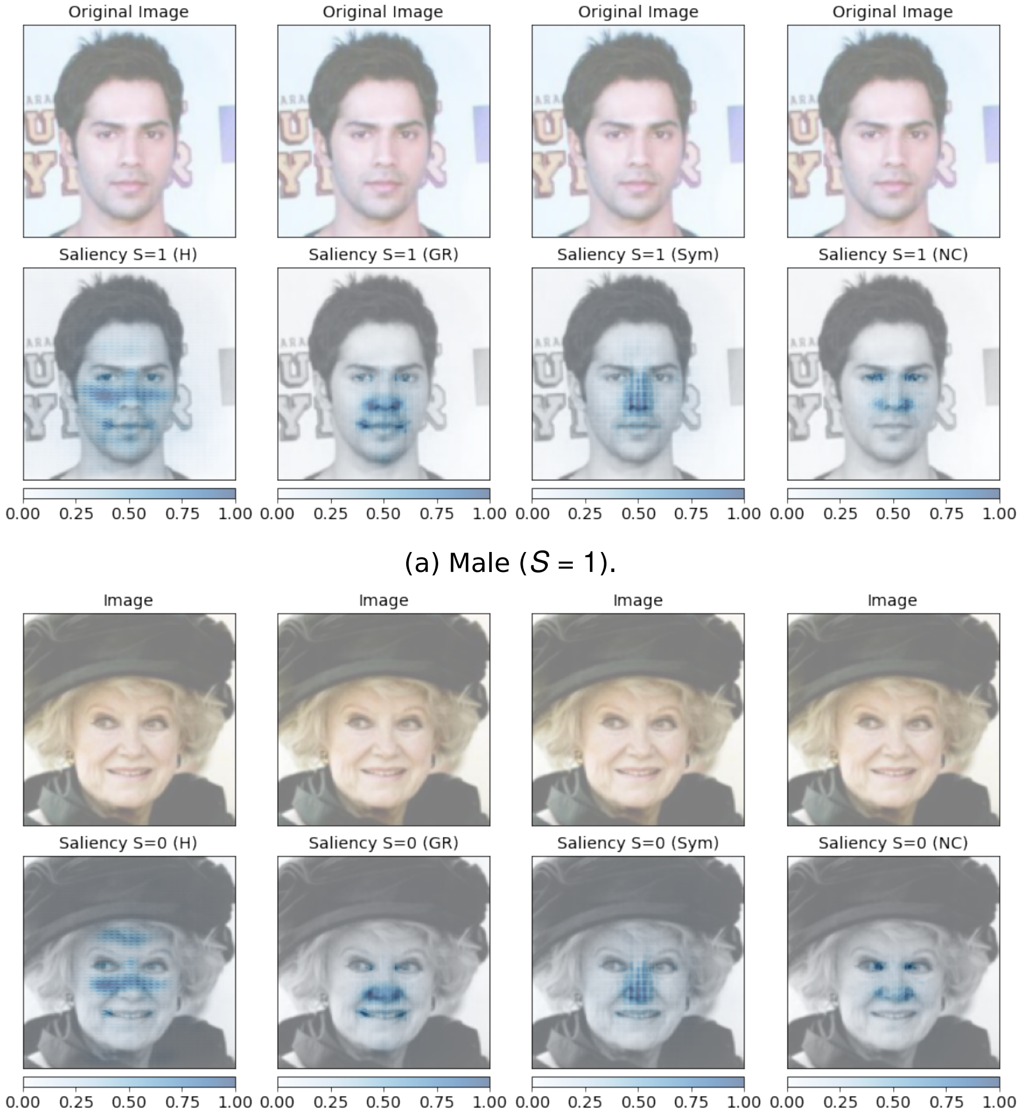


Figure 4.2: Average saliency for the best individual models for the *gender* sensitive attribute. From left to right: model trained on (1) subjective human-based (*H*) labels, and models trained on objective (2) golden ratio (*GR*), (3) symmetry (*Sym*) and (4) neoclassical canons (*NC*), respectively.

The model trained on golden ratio annotations attends to the nose, and edges of the mouth and eyes. In contrast, the model trained on facial symmetry almost entirely focuses on the nose. Finally, the model trained on neoclassical canons annotations attends to the edges of the eyes, nostrils and slightly the border of the mouth. However, none of these models focused on non-informative parts of the face, such as the one dominated



(a) Male ($S = 1$).

(b) Female ($S = 0$).

Figure 4.3: Average saliency for the best individual models for the *gender* sensitive attribute. From left to right: model trained on (1) subjective human-based (H) labels, and models trained on objective (2) golden ratio (GR), (3) symmetry (Sym) and (4) neoclassical canons (NC), respectively.

by skin, *e.g.*, cheeks and forehead, as is the case for the model trained on subjective annotations.

Next, we compute the average region per sensitive attribute. Our goal was to analyze whether the models attended to different regions according to the gender. In Figure 4.3, we can visualize the average saliency regions per gender. We can see that the regions used for each of the models trained on objective annotations, *i.e.*, golden ratio (GR), neoclassical canon (NC) and facial symmetry (Sym), are similar across both genders. In other words, the gender does not seem to play a huge role on which region these models will attend to the most. However, the model trained on subjective annotations (H) have an overall different salient region for male and for female. Specifically, for male, the model seems to use the facial region between the eyes, and the region where usually male people grow their beards. This is not the case for female faces, where the model focuses mostly on the forehead and upper cheeks and nose.

Thus, in summary, in this section, we observed that each model uses a different region. This consolidates previous quantitative results that pointed out that each model has a different behavior regarding accuracy and fairness metric. Moreover, we observed that, in general, models trained on objective annotations attends more to regions supported by previous work that used the mathematical notions for classifying attractiveness (Schmid et al., 2008), while the one trained on subjective labels usually uses a more spread region.

4.2 Facial Expression

In this section we describe the results for the facial expression recognition task.

4.2.1 Dataset Distribution

As it is the case for the attractiveness trait, our goal when generating different t choices for the ObjLCS$_t$ algorithm was to study the effects of different data distributions in the model's behavior. Thus, in other words, in this section, we aim to analyze whether slightly altering the dataset distribution heavily affects the models' behavior. We highlight the fact that the lower the t, the more spread instances are according to all seven facial expressions. In Table 4.9 we show the distribution of the dataset for each t for both the happiness and anger attributes, as well as the distribution of the dataset when annotating it using ObjBase ($\mathcal{D}_{ObjBase}$). The information is scattered across the sensitive attribute of gender expression, *i.e.* male ($S = 1$) and female ($S = 0$). We also added the original distribution of both training datasets, $\mathcal{D}_H$ in ExpW (Happiness) and AffectNet (Anger), and the CFD test splits.

We first note that, due to the the fact that we are dealing with a binary classification (Chen and Joo, 2021), all five expressions that are not considered as happy in the happiness classification, or anger in the anger classification, are annotated as unhappy/non-angry. Thus, as it has been observed by previous work in the FER field (Li and Deng, 2018), we obtain an imbalanced training dataset for subjective and objective annotations. We can visualize that, for the happiness attribute, the base algorithm (ObjBase) for generating the objective labels is the closest one to the distribution of the the labels provided by humans. This was also observed in Section 3.2.1, before we binarized the labels into

Dataset	$\mathcal{D}$	$Y = 1$	$S = 1$	$S = 0$	$Y = 0$	$S = 1$	$S = 0$
ExpW (Happiness)	$\mathcal{D}_{ObjBase}$	25.4%	64.8%	35.2%	74.6%	70.5%	29.5%
	$\mathcal{D}_{ObjLCS_{0.3}}$	15.0%	62.0%	38.0%	85.0%	70.3%	29.7%
	$\mathcal{D}_{ObjLCS_{0.4}}$	14.5%	61.9%	38.1%	85.5%	70.2%	29.8%
	$\mathcal{D}_{ObjLCS_{0.5}}$	13.5%	62.1%	37.9%	86.5%	70.1%	29.9%
	$\mathcal{D}_{H}$	33.1%	63.2%	36.8%	66.9%	71.9%	28.1%
CFD (Happiness)	$\mathcal{D}_{H}$	36.3%	50.0%	50.0%	63.7%	50.0%	50.0%
AffectNet (Anger)	$\mathcal{D}_{ObjBase}$	0.3%	63.4%	36.6%	99.7%	52.1%	47.9%
	$\mathcal{D}_{ObjLCS_{0.3}}$	7.9%	65.2%	34.8%	92.1%	51.1%	48.9%
	$\mathcal{D}_{ObjLCS_{0.4}}$	6.8%	66.0%	34.0%	93.2%	51.2%	48.8%
	$\mathcal{D}_{ObjLCS_{0.5}}$	5.6%	66.7%	33.3%	94.4%	51.3%	48.7%
	$\mathcal{D}_{H}$	5.8%	82.9%	17.1%	94.2%	50.3%	49.7%
CFD (Anger)	$\mathcal{D}_{H}$	17.5%	50.0%	50.0%	82.5%	50.0%	50.0%

Table 4.9: Dataset distribution for both happy and angry facial expressions. For the happy attribute, we used the ExpW Zhang et al. (2015, 2018b) as training dataset, while for the angry attribute we used the AffectNet Mollahosseini et al. (2017) dataset. However, following Chen and Joo (2021), during training, we sample $20k$ instances and average the results over 5 runs. For both happy/angry ($Y = 1$) and unhappy/not angry ($Y = 0$), we also add their distribution with respect to gender, *i.e.*, male ($S = 1$) and female ($S = 0$). $\mathcal{D}_{ObjBase}$ and $\mathcal{D}_{ObjLCS_t}$ correspond to the different dataset distributions of both objective definitions for annotating the facial expressions using AUs, and $\mathcal{D}_H$ represent the distribution of the labels provided by humans. CFD (Happiness) and CFD (Anger) correspond to the test set used to evaluate models trained on happy and angry attributes, respectively Chen and Joo (2021).

happy/unhappy, and it happens for both the main attribute (happiness), as well as for the sensitive attribute (gender).

However, this is not the case when it comes to the anger attribute, which obtains 0.3% of the annotations as containing instances labelled as angry. This might happen due to the fact that ObjBase is very strict when it comes to labeling the facial expressions, *i.e.*, it annotates the facial expression only if all the AUs that represent that expression (Ekman, 1993) are detected. According to previous wok on the literature (Ekman, 1993), happiness requires that only two AUs to be active (6 and 12), while anger requires that four AUs be active (4, 5, 7 and 23) to be labeled as an happy/angry instance, respectively. Thus, ObjBase algorithm results in fewer instances annotated as containing an angry expression. This hurts the diversity of annotations, as it was described in Section 3.2.1.

For the ObjLCS$_t$, since it generates labels that are more spread across all the facial expressions, it provides a more imbalanced dataset with respect to the binary happiness attribute. On the other hand, since an angry face is composed of multiple AUs, this algorithm provides a more balanced distribution for the anger dataset. In other words, the dataset distributions that ObjLCS$_t$ produces for all t values are more balanced for the anger attribute, while for the happiness attribute it is slightly less balanced, both compared to their corresponding distribution based on ObjBase. This happens due to the fact that ObjLCS$_t$ considers more the intensity of each AU, even the ones that were not necessarily detected.

Regarding the distribution of female and male with respect to both happy and unhappy instances, for all the annotation procedures, the proportion is similar. For the anger attribute, the distribution of gender expressions is more balanced for non-angry ($Y = 0$) instances. We highlight the fact that the distribution of the test set was purposely modified by Chen and Joo (2021) such as the allocation of happy/angry and unhappy/non-angry images between male and female is the same, *i.e.*, the distribution of female/male is 50%. Nonetheless, the distribution of happy and unhappy, and angry and non-angry instances in the test datasets of CFD (Happiness) and CFD (Anger), respectively, is also imbalanced, *i.e.*, 36.3% happy ×63.7% unhappy and 17.5% angry ×82.5% non-angry. Since the CFD splits were used solely for evaluation purposes, we do not modify them.

4.2.2 Individual Models

Similarly to the atttactive attribute, our goal in this sections is to verify whether training models on objective labels of the ExpW and AffectNet datasets actually produce improvements in the fairness metric for the CFD dataset (Chen and Joo, 2021), compared to the ones trained on subjective labels. Table 4.10 depicts the average results across five different runs of individually training the models on different definitions of facial expression for the happiness and anger attributes. We trained models using the algorithms described in Section 3.2.1, named ObjBase and ObjLCS$_t$. We tested all thresholds t shown in Figure 3.5. We also added a row named H, which depicts the average result of the models trained on human-based ExpW and AffectNet annotations.

As is the case for the attractive attribute, in general, we notice a trade-off between accuracy and fairness ($\Delta Disc$). For the happiness attribute, this is mainly the case for the models trained using the ObjLCS$_t$ algorithm since they obtain a lower $\Delta Disc$ value than the models trained with the ObjBase labels and human-based labels. Even though all the ObjLCS$_t$ obtained similar $\Delta Disc$ values, the one which in the best (lowest) $\Delta Disc$ metric is the $t = 0.3$. This might be related to the fact that the annotations produced by this threshold are more more spread across all expressions. Moreover, for the happiness attribute, the algorithm ObjBase obtains competitive accuracy results with the models trained on human labels (H), with the expense of having a relatively similar and high fairness measure. This is expected since in the previous section (Section 4.2.1) we observed that the ObjBase algorithm produced a distribution similar to the one of the human-based labels.

For the anger attribute, we can infer an opposite behavior, in which the model that obtained the lowest $\Delta Disc$ is trained with annotations from the ObjBase algorithm.

Attribute	Annotation Algorithm	Accuracy	$\Delta Disc$
Happiness	ObjBase	0.926 ± 0.007	0.052 ± 0.019
	ObjLCS$_{0.3}$	0.826 ± 0.027	**0.009** ± 0.021
	ObjLCS$_{0.4}$	0.829 ± 0.013	0.013 ± 0.022
	ObjLCS$_{0.5}$	0.816 ± 0.017	0.013 ± 0.027
	H	0.935 ± 0.009	0.046 ± 0.025
Anger	ObjBase	0.825 ± 0.000	**0.028** ± 0.018
	ObjLCS$_{0.3}$	0.825 ± 0.002	0.043 ± 0.039
	ObjLCS$_{0.4}$	0.824 ± 0.001	0.077 ± 0.081
	ObjLCS$_{0.5}$	0.825 ± 0.001	0.064 ± 0.048
	H	0.855 ± 0.046	0.055 ± 0.045

Table 4.10: Results of models trained on different happiness and anger notions. Each model was evaluated according to its attribute (anger and happiness) on CFD test set. We show the average results across five different seeds for the sensitive attribute *gender* (Chen and Joo, 2021). ObjBase and ObjLCS$_t$ correspond to the different annotation algorithms (Annotation Algorithm) using AUs, and H represent the original labels, based on human perception. We highlighted the best and underlined the worst average $\Delta Dist$ results for the models trained on objective annotations.

However, we note that the standard deviation is higher in $\Delta Disc$ for this attribute. We suspect that this behavior is more dependent upon the seed since the anger attribute possesses a more imbalanced data distribution than the happiness attribute. The model trained on $ObjLCS_{0.5}$ obtains similar results to the one obtained with human labels (H), again following the trend that the model which has the closer data distribution to the original labels obtain similar $\Delta Disc$.

In the case of facial expression recognition, we observe that the drop in accuracy was not as severe as the one observed for the attractiveness classification. Nonetheless, from the results shown above, we can conclude that individually training models on the objective notions of facial expression tends to improve the fairness metrics.

4.2.3 Ensemble Model

The ensembles in this section are produced by a weighted combination of three models, *i.e.*, H, ObjBase and ObjLCSt, each trained on a different definition of facial expression. We followed the same procedure as in the attractive attribute, and randomly chose one seed for all models. Figure 4.4 shows the result of several weighting values per model. The plots on the left (Figure 4.4a and Figure 4.4c) correspond to the ensemble results for the selected best individual models, *i.e.*, models which obtained best results with respect to $\Delta Dist$ (H, ObjBase and ObjLCS$_{0.3}$, for both attributes), while the one on the right (Figure 4.4b and Figure 4.4d) depicts the result for the worst models (H, ObjBase and ObjLCS$_{0.5}$ for the happiness attribute, and H, ObjBase and ObjLCS$_{0.4}$ for the anger attribute).

We show the result of each ensemble with respect to accuracy and $\Delta Disc$ in Figure 4.4. Each blue dot illustrates the result of one ensemble model, and the individual models, *i.e.*, models that use a single happiness/anger definition to compose its final decision, are shown as crosses: the model trained on human-based annotations (H) is shown in red, and the ones trained only with mathematical concepts, such as ObjBase and ObjLCSt, are shown in green and orange, respectively. We varied the weight of each individual (base) model from 0 to 1 with steps 0.05. Finally, the gray dots represent the Pareto analysis.

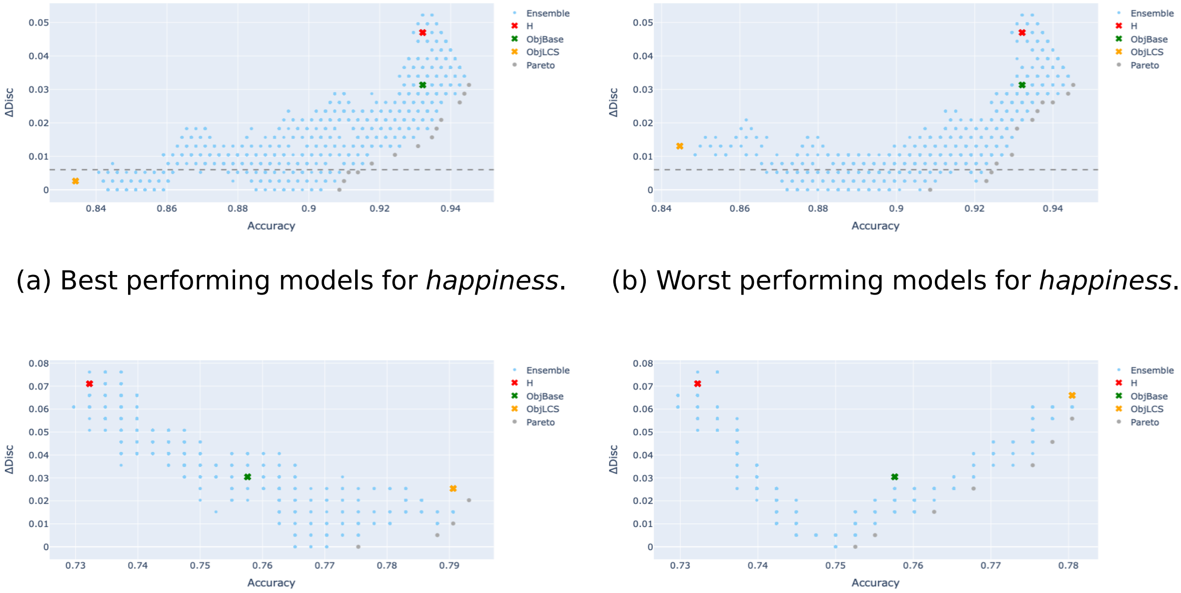
We observe that the plots for the happiness attribute (Figure 4.4a and Figure 4.4b) show comparable curve results, suggesting again that individually combining the best and worst models into an ensemble have approximately the same result regarding overall accuracy and $\Delta Disc$. We even visualize that the ensembles that combine the worst performing models have more models with $\Delta Disc = 0$. This is possible due to the fact that this is a less strict metric, relying exclusively in keeping the proportion between predictions of both sensitive groups balanced, *i.e.*, it does not use any information from the annotated labels. Moreover, this reinforces the idea that the choice of t does not have a huge impact on the final result.

As is the case for the attractive attribute, the final result is heavily dependent on the weights each base model has in the final ensemble. This can be directly inferred from how scattered the ensemble models are. Thus, from Figure 4.4a and Figure 4.4c we can see that there is a wide range of possible models. For instance, from the horizontal gray line, fixed at $\Delta Disc = 0.006$ (Chen and Joo, 2021), it is possible to obtain an overall accuracy of more than 0.9.

The results for the anger attribute, shown in Figure 4.4c and Figure 4.4d, show a slightly different pattern. The best and worst performing models express different curves. This happens specially due to the ObjLCS$_t$ model. This is complementary to what we have observed in the previous section for the individual models, where the individual models obtained a higher standard deviation. In other words, the individual results vary a lot depending on the seed. We hypothesize this is due to the fact that the dataset is much more imbalanced for the anger attribute than for the happiness attribute.

Nonetheless, as it is the case for the happiness classification, we obtain models that reach low levels of $\Delta Disc$. Some even achieve $\Delta Disc = 0$. Moreover, the range of accuracy values is lower for the anger (which varies from ≈ 0.73 to ≈ 0.79) than for the happiness attribute (which varies from ≈ 0.84 to ≈ 0.94). The opposite happens for the fairness metric.

Thus, in summary, we observed in this section that it is possible to obtain a wide range of possible results when combining the individual models through a weighted average. More importantly, we showed that it is possible to select models that achieve high accuracy and low unfairness. As is the case for the attractiveness classification task, the



(a) Best performing models for *happiness*.

(b) Worst performing models for *happiness*.

(c) Best performing models for *anger*.

(d) Worst performing models for *anger*.

Figure 4.4: Results of different weighting values to compose the final ensemble for the facial expression attributes happiness and anger. Plots on the first and second column correspond to the best and worst individual models with respect to $\Delta Disc$ in Table 4.10, respectively. We show the result of each ensemble with respect to accuracy (x axis) and $\Delta Disc$ (y axis) for the sensitive attribute *gender*. Light blue dots represent the *ensemble* models, *i.e.*, combining different annotations; red, green, and orange crosses indicate the model trained with only human-based annotations (H), base (ObjBase) and LCS (ObjLCSt) objective annotations, respectively; finally, the gray dots represent the Pareto analysis.

decision upon which ensemble to choose from will depend deeply on the downstream application.

4.2.4 Comparison with Prior Work

In this section, we compare our method with previous debiasing approaches for the FER system. The goal of this section is to analyze whether our method obtains a better trade-off between accuracy and fairness compared to previous related work. In order to choose some ensembles over all possible combinations, we follow previous work (Chen and Joo, 2021) and calculate the average and standard deviation ($\pm$) over some of the runs. Since the combination of all three models for each of the five seeds would require computing $5^3 = 3125$ results, we randomly selected 2 of the 5 seeds ($2^3 = 8$ combinations). Previous work do not provide the overall accuracy (Chen and Joo, 2021), thus we decided to run all the debiasing methods and report them according to the code provided by Chen and Joo (2021).

Method	Accuracy	$\Delta Disc$
Baseline (Chen and Joo, 2021)	-	0.059 ± 0.035
Baseline (our H)	$0.935 \pm 0.009^*$	$0.046 \pm 0.025^*$
Uniform Confusion (Alvi et al., 2018)	$0.934 \pm 0.008^*$	0.046 ± 0.008
Gradient Projection (Zhang et al., 2018a)	$0.842 \pm 0.107^*$	0.036 ± 0.014
Domain Discriminative (Wang et al., 2020)	$0.931 \pm 0.013^*$	0.076 ± 0.024
Domain Independent (Wang et al., 2020)	$0.920 \pm 0.021^*$	0.029 ± 0.015
AUC-FER (Chen and Joo, 2021)	$\underline{0.900} \pm 0.009^*$	$\underline{0.006} \pm 0.020$
$Ours_1$ ($H = 0.25; ObjBase = 0.05; ObjLCS_{0.3} = 0.35$)	$\underline{0.894} \pm 0.008$	$\mathbf{0.005} \pm 0.005$
$Ours_2$ ($H = 0.50; ObjBase = 0.05; ObjLCS_{0.3} = 0.65$)	$\underline{0.895} \pm 0.009$	$\mathbf{0.005} \pm 0.007$
$Ours_9$ ($H = 0.10; ObjBase = 0.55; ObjLCS_{0.3} = 0.55$)	0.907 ± 0.010	0.009 ± 0.006
$Ours_{10}$ ($H = 0.05; ObjBase = 0.50; ObjLCS_{0.3} = 0.40$)	0.915 ± 0.006	0.010 ± 0.005

Table 4.11: Accuracy and fairness scores ($\Delta Disc$) for previous debiasing approaches on the *happiness* attribute. Following previous work (Chen and Joo, 2021), for our models, we report the average and standard deviation (±) across two seeds for each model. The symbol * represents the values we obtained when reproducing previous work, according to the code provided by Chen and Joo (2021). We highlighted the best and underlined similar average $\Delta Dist$ results.

In order to choose some of all the possible ensembles, and considering that none of the best performing models occur in all eight combinations at the same time, we first run the Pareto analysis on each combination, and select the models contained on each one of them individually. Then we calculate the average and standard deviation for all models in the eight Pareto boundaries, and sort them regarding $\Delta Disc$ (the lower the fairer the model is). We additionally remove ensemble models that obtain the same accuracy and fairness metric, *i.e.*, ensembles with duplicated results.

Table 4.11 depicts the results for the happiness attribute. We selected four models ($Ours_1$, $Ours_2$, $Ours_9$, $Ours_{10}$), two of them performed the best regarding $\Delta Disc$ ($Ours_1$, $Ours_2$), and two of them which achieved the best overall accuracy in the top-10 best ensemble models ($Ours_9$, $Ours_{10}$). We compare our method with previous debiasing approaches for the attractive attribute (Alvi et al., 2018; Zhang et al., 2018a; Wang et al., 2020; Chen and Joo, 2021) using the sensitive attribute *gender*. For completeness, we show all the 10 best performing ensemble models with respect to $\Delta Disc$ in Table 4.12.

We first note that the top-2 performing models regarding $\Delta Disc$ obtained a metric that is lower than the previous state-of-the-art, while holding a competitive overall accuracy (≈ 0.9). Nonetheless, the models that obtained the best overall accuracy in the top-10 performing ensembles also have a lower $\Delta Disc$ than most of previous work, except for the work of Chen and Joo (2021). Differently from the best performing models on the attractiveness attribute, in Table 4.12 we observe a diverse set of weights for each individual model that composes the final ensembles, *i.e.*, no single model obtains the lowest

	Weights			Evaluation Metrics	
	H	*ObjBase*	$ObjLCS_{0.3}$	Accuracy	Δ*Disc*
1	0.25	0.05	0.35	0.894 ± 0.008	0.005 ± 0.005
2	0.50	0.05	0.65	0.895 ± 0.009	0.005 ± 0.007
3	0.00	0.35	0.30	0.905 ± 0.009	0.006 ± 0.007
4	0.00	1.00	0.85	0.905 ± 0.009	0.006 ± 0.008
5	0.30	0.00	0.35	0.898 ± 0.008	0.007 ± 0.010
6	0.35	0.00	0.40	0.899 ± 0.009	0.007 ± 0.008
7	0.10	0.50	0.55	0.901 ± 0.009	0.007 ± 0.006
8	0.60	0.00	0.65	0.903 ± 0.010	0.008 ± 0.008
9	0.10	0.55	0.55	0.907 ± 0.010	0.009 ± 0.006
10	0.05	0.50	0.40	0.915 ± 0.006	0.010 ± 0.005

Table 4.12: Weights and metrics of the top-10 performing ensembles sorted by Δ*Disc* (the lower the fairer the model is) for the *happiness* attribute. We show overall error rate and Δ*Disc*. Following previous work (Chen and Joo, 2021), we report the average and standard deviation (±) across some seeds for each model. Specifically, we randomly chose 2 random seeds for each model that compose the final ensemble, resulting in $2^3 = 8$ combinations. For selecting the top performing models, we then ran the Pareto analysis over each combination. Since none of the top performing models appear in all eight combinations (*i.e.*, no model appear in the intersection of the best performing models for all eight combinations), we selected the models that appear in each of the eight Pareto boundaries individually. We then calculate the average and standard deviation across the chosen seeds for the selected top performing models, sorting them by Δ*Disc*.

influence over the final ensemble. Nonetheless, we observe that the model trained on $ObjLCS_{0.3}$ always obtains a weight close to *either H* or *ObjBase*.

However, the models trained on *H* and *ObjBase* labels never obtain a similar weight across all the ensembles. We hypothesize that this is due to the fact that these models obtain similar results for both metrics (and behaviors, as it will be visually clearer in Section 4.2.5), *i.e*. both obtained high accuracy *and* high Δ*Disc*. Thus, increasing the weight of either of them, combined with the model that had the best (*i.e*., lowest) individual result regarding the fairness metric ($ObjLCS_{0.3}$), optimizes for both overall accuracy and fairness. We note that two of the four selected results have the lowest Δ*Disc* ($Ours_1$ and $Ours_2$) of all the previous methods. Both the ensembles that obtains a similar but slightly higher Δ*Disc*, as the work of Chen and Joo (2021) ($Ours_9$ and $Ours_{10}$) also have a slightly higher accuracy than this previous work.

In Table 4.13 we report the results for the anger attribute. Previous work (Chen and Joo, 2021) used an additional balancing procedure for this attribute. Thus, we reproduce their results, but we maintain the same procedure as for the happiness attribute, *i.e*., without any data balancing. We also followed the same methodology for selecting the best performing models, *i.e*., we ran the Pareto analysis on the eight possible combinations (2 seeds for each of the 3 models), calculated the average and standard deviation of all metrics and sorted them according to Δ*Disc*. In Table 4.13, we show the four best per-

Method	Accuracy	$\Delta Disc$
Baseline (our H)	$0.902 \pm 0.028^*$	$0.055 \pm 0.045^*$
Domain Discriminative (Wang et al., 2020)	$0.903 \pm 0.023^*$	$0.059 \pm 0.045^*$
Domain Independent (Wang et al., 2020)	$0.877 \pm 0.039^*$	$0.092 \pm 0.109^*$
AUC-FER (Chen and Joo, 2021)	$0.840 \pm 0.019^*$	$0.060 \pm 0.032^*$
$Ours_1$ ($H = 0.00; ObjBase = 0.65; ObjLCS_{0.3} = 0.05$)	0.782 ± 0.004	**0.010** ± 0.007
$Ours_2$ ($H = 0.00; ObjBase = 0.80; ObjLCS_{0.3} = 0.05$)	0.780 ± 0.004	**0.011** ± 0.010
$Ours_9$ ($H = 0.00; ObjBase = 0.25; ObjLCS_{0.3} = 0.40$)	0.789 ± 0.011	**0.023** ± 0.011
$Ours_9$ ($H = 0.00; ObjBase = 0.15; ObjLCS_{0.3} = 0.25$)	0.789 ± 0.001	**0.024** ± 0.010

Table 4.13: Accuracy and fairness scores ($\Delta Disc$) for previous debiasing approaches on the *anger* attribute. Following previous work (Chen and Joo, 2021), for our models, we report the average and standard deviation (±) across two seeds for each model. The symbol * represents the values we obtained when reproducing previous work, according to the code provided by Chen and Joo (2021). Since previous work use an additional balancing algorithm for this attribute, we only report the the results we obtained when reproducing them. We highlighted the best average results.

	Weights			Evaluation Metrics	
	H	$ObjBase$	$ObjLCS_{0.3}$	Accuracy	$\Delta Disc$
1	0.00	0.65	0.05	0.782 ± 0.004	0.010 ± 0.007
2	0.00	0.80	0.05	0.780 ± 0.004	0.011 ± 0.010
3	0.05	0.45	0.30	0.771 ± 0.009	0.013 ± 0.012
4	0.05	0.40	0.60	0.778 ± 0.008	0.014 ± 0.010
5	0.05	0.75	0.55	0.777 ± 0.008	0.014 ± 0.012
6	0.05	0.00	0.50	0.778 ± 0.009	0.018 ± 0.006
7	0.05	0.05	0.30	0.773 ± 0.008	0.018 ± 0.015
8	0.00	0.25	0.05	0.784 ± 0.008	0.022 ± 0.008
9	0.00	0.25	0.40	0.789 ± 0.011	0.023 ± 0.011
10	0.00	0.15	0.25	0.789 ± 0.001	0.024 ± 0.010

Table 4.14: Weights and metrics of the top-10 performing ensembles sorted by $\Delta Disc$ (the lower the fairer the model is) for the *anger* attribute. We show overall error rate and $\Delta Disc$. Following previous work (Chen and Joo, 2021), we report the average and standard deviation (±) across some seeds for each model. Specifically, we randomly chose 2 random seeds for each model that compose the final ensemble, resulting in $2^3 = 8$ combinations. For selecting the top performing models, we then ran the Pareto analysis over each combination. Since none of the top performing models appear in all eight combinations (*i.e.*, no model appear in the intersection of the best performing models for all eight combinations), we selected the models that appear in each of the eight Pareto boundaries individually. We then calculate the average and standard deviation across the chosen seeds for the selected top performing models, sorting them by $\Delta Disc$.

forming models ($Ours_1$, $Ours_2$, $Ours_9$, $Ours_{10}$), two of them performed the best regarding $\Delta Disc$ ($Ours_1$, $Ours_2$), and two of them which achieved the best overall accuracy in the top-10 best ensemble models ($Ours_9$, $Ours_{10}$). We show the complete list of all the 10 best performing ensemble models with respect to $\Delta Disc$ for the anger attribute in Table 4.14.

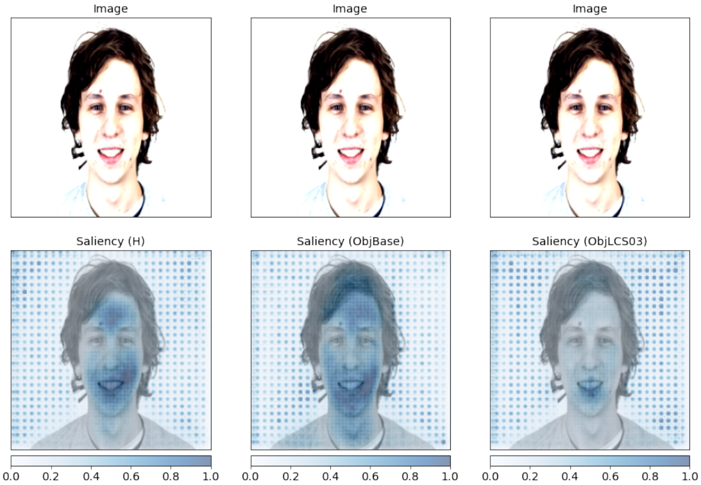


Figure 4.5: Average saliency for the best individual models for the *happiness attribute*. From left to right: model trained on subjective human-based (H) labels, and models trained on objective labels generated by the ObjBase and ObjLCS$_{0.3}$ algorithms, respectively.

We first note from the table that all of our models obtained a lower fairness metric, *i.e.*, had a less discriminatory behavior according to $\Delta Disc$. Specifically, we decrease six times compared to the best state-of-the-art approach (*i.e.*, AUC-FER (Chen and Joo, 2021)) which resulted in $\Delta Disc$ = 0.060. However, for the anger attribute, this came with a higher cost regarding accuracy. While the baseline obtains near 0.90 accuracy, our approach obtains $\approx$ 0.79.

We can also observe that, for both anger and happiness classification, our ensembles obtained a lower standard deviation compared to previous methods. This indicates that our models behave in a more stable manner. Thus, in summary, in this section, as is the case for the attractive attribute, we obtained a satisfactory trade-off between a given fairness metric ($\Delta Disc$) and accuracy for both FER systems. We demonstrated that our method is simultaneously useful (*i.e.*, obtains a high accuracy) and effective at mitigating biases.

4.2.5 Model Interpretation

In this section, we describe the results when applying an explainability method to the individual models that compose the final ensemble. For this analysis, we follow the same procedure as in the attractive attribute and use the saliency method. Thus, once again, we begin by computing the average region that each model attended for all the instances in the dataset. We randomly sampled one image from the test set to plot the average saliency.

The results for the happiness attribute can be visualized in Figure 4.5. We first note that all the models heavily attend the mouth region to make its final decision. This could be attributed to the fact that our models are trained on a binary classification task for detecting happy/unhappy instances, and smiling has always been viewed as an easily identifiable indicator of an individual's happiness (Moore et al., 2017).

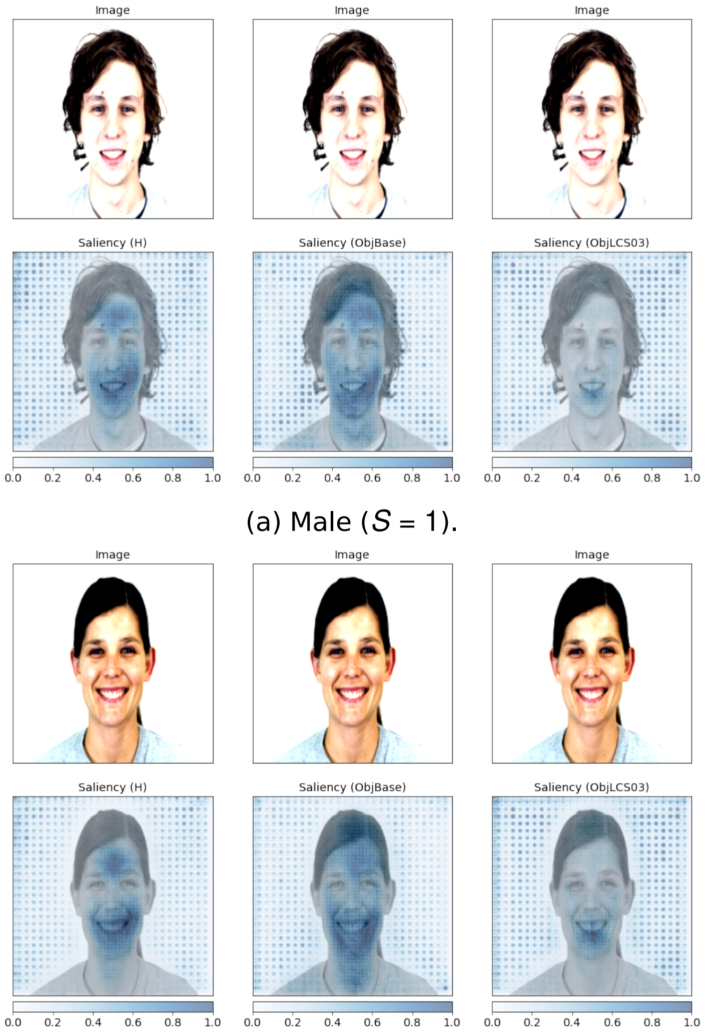
Additionally, we note that both the model trained on subjective human-based annotations and the one trained with the annotations generated by the ObjBase algorithm attended a bigger and more spread region across the face. Both models use mainly the forehead, mouth and cheeks region to classify expressions in images. However, the model trained on annotations generated by the $\text{ObjLCS}_{0.3}$ contains a smaller and more concentrated region. Specifically, this model concentrates its attention mostly in the mouth region. This might be one of the reasons why this model obtained a lower $\Delta Disc$ in Section 4.10.

Next, we compute the average region per sensitive attribute. Our goal is to verify whether, on average, the models pay attention to different regions according to the gender. In Figure 4.6, we can visualize the average saliency regions per gender. We can see that across male and female regions, the models trained on subjective and objective annotations use, in general, similar regions to make the final decision. In other words, the blue regions are similar across genders for the same model.

We also applied the saliency method to the individual models that compose the ensemble for the anger attribute. In Figure 4.7 we show the average region that each model attended for all the instances in the dataset. Again we sampled one image of the whole dataset just to make the visualization easier. For this attribute, we notice that the baseline attends to the forehead and some regions near the mouth. This corresponds to the regions described by the AUs that compose the anger expression, *i.e.*, AU04 (Brow Lowerer), AU05 (Upper Lid Raiser), AU07 (Lid Tightener) and AU23 (Lip Tightener). Similarly, the model trained on ObjLCS algorithm also attends to forehead, though it pays less attention to the mouth region. Finally, the model trained on ObjBase attends to some regions of the eyes, but it contains a more spread area than the other models.

Next, as it was done for the happiness attribute, we compute the average region per sensitive attribute ($S = 0$ and $S = 1$). In Figure 4.8, we can visualize the average saliency regions per gender. Again we notice that, for the same model, the regions across both male and female is similar. However, for some models, *e.g.*, ObjBase, the intensity slightly alters across male and female instances. Nonetheless, we can infer that the models use the same regions to make its final decision, regardless of the gender expression.

Thus, in summary, in this section, we observed that each model uses a different region. This consolidates previous quantitative results that pointed out that each model has a different behavior regarding accuracy and fairness metric. Moreover, we observed that, in general, models trained on objective annotations attends more to regions sup-



(a) Male ($S = 1$).

(b) Female ($S = 0$).

Figure 4.6: Average saliency for the best individual models for the happiness attribute, scattered along gender expressions. From left to right: model trained on subjective human-based (H) labels, and models trained on objective labels generated by the ObjBase and ObjLCS$_{0.3}$ algorithms, respectively.

ported by previous work that developed FACS (Ekman, 1993), while the ones trained on subjective labels usually use a more spread region.

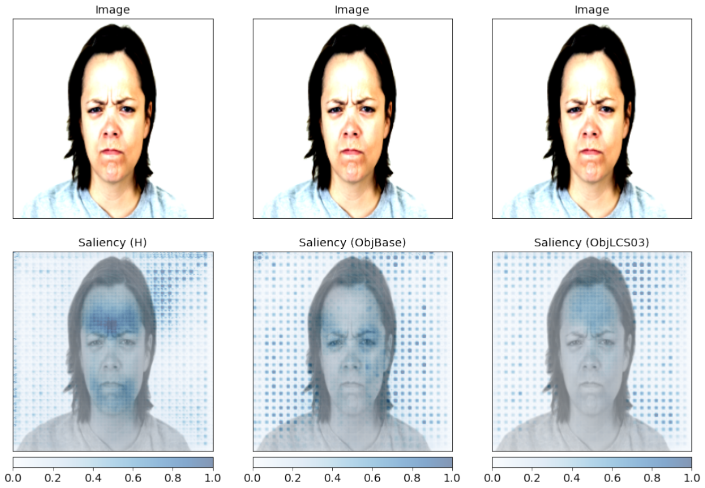
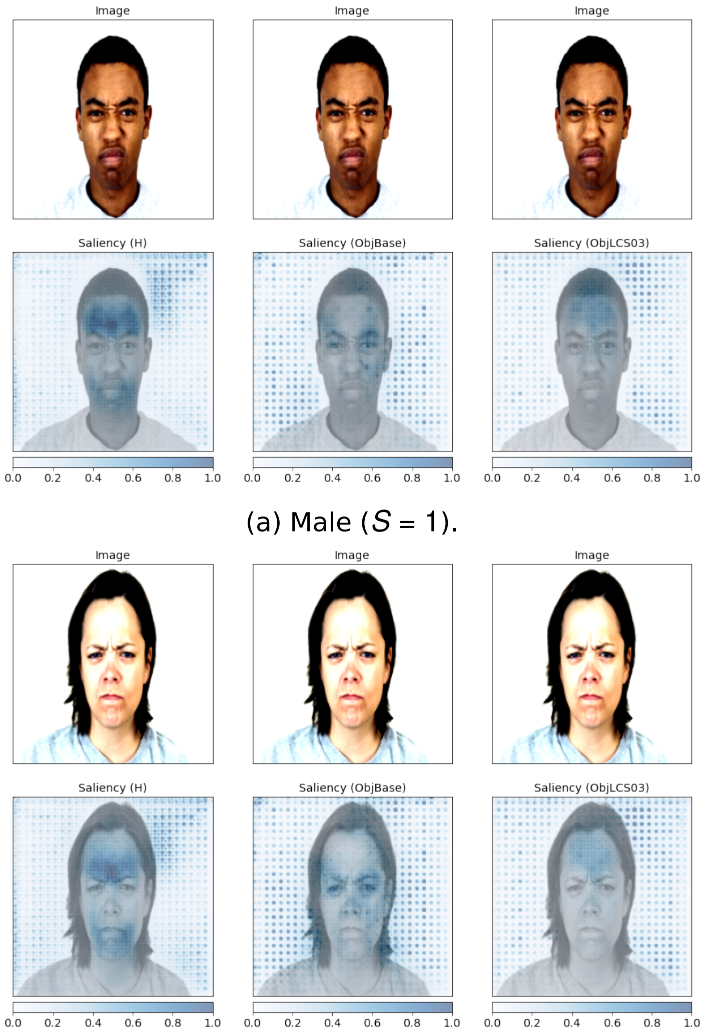


Figure 4.7: Average saliency for the best individual models for the *anger attribute*. From left to right: model trained on subjective human-based (H) labels, and models trained on objective labels generated by the ObjBase and ObjLCS$_{0.3}$ algorithms, respectively.



(a) Male (S = 1).

(b) Female (S = 0).

Figure 4.8: Average saliency for the best individual models for the anger attribute, scattered along gender expressions. From left to right: model trained on subjective human-based (H) labels, and models trained on objective labels generated by the *ObjBase* and $ObjLCS_{0.3}$ algorithms, respectively.

5. FINAL REMARKS

In this section, we first introduce the ethical considerations of our work. Next, we make the final considerations.

5.1 Ethical Considerations

The technique proposed in this paper can be applied to mitigate unintended and undesirable biases in some facial analysis systems. While the idea behind our proposed method is important and can be broadly applied to many other domains, it is not sufficient. Rather, as described in Denton et al. (2019) it must be part of a larger, socially contextualized project to critically assess ethical concerns relating to facial analysis technology. This project must include addressing questions of whether and when to deploy technologies, frameworks for democratic control and accountability, and design practices which emphasize autonomy, inclusion, and privacy.

Regarding dataset choice, in this work we use CelebA dataset (Liu et al., 2015) for the attractive attribute, and Expw (Zhang et al., 2015, 2018b), AffectNet (Mollahosseini et al., 2017) and CFD (Ma et al., 2015) for the facial expression attribute. As all of these datasets contain public domain images, it avoids the issues of some other public domain datasets of face images (*e.g.*, Klare et al. (2012)). Moreover, as mentioned before, all datasets used in this work are well-known benchmarks in the ML community, with previous proposed methods on mitigating fairness-related issues (Ramaswamy et al., 2021; Sattigeri et al., 2019; Quadrianto et al., 2019; Chen and Joo, 2021). The attributes within the CelebA dataset are reported as binary categories, and for the ExpW, AffectNet and CFD datasets we followed the procedure described on the work of Chen and Joo (2021) to binarize the facial expressions into happy/unhappy and angry/non-angry. We note that in many cases this binary categorization does not reflect the real human diversity of attributes. This is perhaps most notable when the attributes are related to continuous factors.

The intent of this work is to demonstrate the utility of reasoning about demographics – specifically in the context of attractiveness and facial expression recognition – in order to do better handle cases where these demographics are used for discriminatory purposes. We highlight the fact that both tasks here were used as a means instead of an end. We do not wish to reinforce any type of prejudice or discrimination based on this measurements, nor motivate inferring these measures for individuals without their consent. Instead, we use these attributes mainly as applications of our proposed method. Additionally, gender is not necessarily the one the person identifies with, rather we considered gender expression, which can be often directly inferred by humans.

Finally, we also note that our method may have other limitations. For instance, we considered datasets collected from in-the-wild images. These images do not have any background, facial orientation or facial emotion pattern. Rather, it contains different background colors, frontal and lateral faces, and several facial expressions. This may present a limitation, since our method, which is based on landmarks and AUs extraction, does not fully work on lateral facial poses. For instance, for the attractive attribute, we tried to mitigate this limitation by removing lateral facial poses from the training and validation set.

5.2 Conclusion

In this work, we studied the fairness issues associated with facial analysis systems. Specifically, we focused our research on two main aspects of facial analysis: attractiveness classification and facial expression recognition, both of which have huge impact on our daily lives. We propose a method that combines different types of annotations, such as the original and subjective human-based labels and the ones we objectively generate based on mathematical definitions of both tasks. Our approach is not only simple, but also intuitive and model-agnostic.

We first demonstrated that the individual models trained solely with objective labels improve the fairness metric, *i.e.*, are less discriminatory, compared to the one trained on subjective labels. We showed that this result is not dependent upon the data distribution of the novel annotations. In other words, our results were not sensitive to the choice of t or δ. We demonstrated that for the attractiveness classification task comes with a cost of reducing the overall accuracy to random choice, *i.e.*, ≈50% in a binary classification setting. However, we also showed that this is not always the case, since for the facial expression recognition the drop in accuracy was not as severe as the one we observed for the attractive attribute. Nonetheless, both reduced the unfair behavior according to their fairness metrics.

We then combined the individual models, each trained on different perspectives of the task at hand, into an ensemble model. Specifically, we use a weighted average combination of all the models. We again showed that our method is robust to the choice of combination between best and worse models, *i.e.*, both curves have a similar shape. From this process we obtained a huge set of possible ensemble models, each one with its own set of weights. We subsequently ran the Pareto efficiency analysis, which aims to optimize for both accuracy and fairness, and selected the models positioned in the Pareto curve as the ones which obtained the best weight combination. We then compare our method to previous work on the field, for both facial analysis tasks, and showed that it is possible to simultaneously maintain a competitive accuracy and reduce the fairness metric.

Finally, we used an explainability technique to understand which region each model attends to the most. We showed that, in general, the models which are trained on objective annotations use a more narrow and specific region, usually around the area of interest for the classification. For instance, in the attractiveness task all models trained on objective labels (golden ratio, symmetry and neoclassical canons) use the nose, eyes and a part of the mouth region to make its prediction. The same applies for the models trained on facial expression recognition, especially the ones trained on the annotations generated by the ObjLCS$_t$ algorithm, which, for instance, focuses on the mouth region for happiness classification. However, the models trained on the subjective labels use a more spread area, accounting for regions that do not provide meaningful information for making the final prediction, *e.g.*, forehead, chin and cheeks. This analysis supports our previous result, in which models trained on objective labels have a different behavior regarding the fairness metric.

Thus, in summary, in this work we demonstrated that by training models on a diverse set of labels we are able to obtain an ensemble that improves the fairness metrics over the baselines, while maintaining competitive accuracy. We also showed that each model that composes the ensemble make its final predictions based on different facial regions and possibly features. To the best of our knowledge, this is the first time a pre-processing debiasing method combines objective (mathematical) labels and subjective (human-based) annotations aiming at reducing fairness issues in ML models. This approach can be extended beyond the tasks explored in this work, and, in general, one can use any objective measures for tasks requiring subjective human labeling within the proposed framework. Although such objective measures may not always be accurate in practice, the belief is that because these measures are often geometrical attributes and are possibly less affected by other attributes of the subjects, they are fairer than the subjective labels in the training data and can thus be used to mitigate fairness issues.

5.3 Future Work

For future work, we plan first on extending our method to more facial analysis systems. We would also like to test our approach under a multi-class setting, in which we relax the constraint of presence ($Y = 1$) or absence ($Y = 0$) of a specific attribute.

Additionally, we would like to study how other types of weighting techniques for ensembles would influence the results we obtained. Thus, instead of a simple weighted average procedure, we could introduce a more sophisticated mechanism for combining the individual models trained on different perspectives. Specifically, we would like to start by testing other debiasing weighted techniques (Bhaskaruni et al., 2019; Kenfack et al., 2021), that was already proposed in the literature, with our annotation methodology.

Finally, another interesting future direction would be to explore which individual models have the strongest influence on debiasing the final ensemble, and why this is the case. For instance, when selecting the best performing ensembles for the attractiveness classification, we observed that some models trained with objective annotations (*e.g.*, golden ratio) did not provide a substantial reduction in the fairness metric, compared to the fairness metric obtained from the model trained with human annotations. This was especially the case when combining this model with the remaining ones in the ensemble, *i.e.*, the weights (influence) that this model had on the best ensembles was considerably close to zero.

5.4 Publications

During the development of the Master's dissertation, some papers have been published:

(1) *How Does Computer Animation Affect Our Perception of Emotions in Video Summarization?* at the International Symposium on Visual Computing, 2020: **Camila Kolling**, Victor Araujo, Rodrigo C. Barros, Soraia Raupp Musse.

(2) *Efficient Counterfactual Debiasing for Visual Question Answering* at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022: **Camila Kolling**, Martin More, Nathan Gavenski, Eduardo Pooch, Otávio Parraga, Rodrigo C. Barros.

(3) *Measuring Representational Robustness of Neural Networks Through Shared Invariances* at the International Conference on Machine Learning (ICML), 2022: Vedant Nanda, Till Speicher, **Camila Kolling**, John P. Dickerson, Krishna P. Gummadi, Adrian Weller.

The first one was developed during the Computer Animation course offered by PUCRS. The second was developed during the first year of Master's, as a side project. The last one resulted from one of the internships conducted during the second year of Master's at the Max-Planck Institute for Software System (MPI-SWS) under the supervision of Prof. Dr. Krishna P. Gummadi. Moreover, the theme of the work presented here has been submitted to and it is currently under review at the IEEE Transactions on Image Processing [1].

[1] Available at https://arxiv.org/abs/2204.06364.

REFERENCES

Alvi, M. Zisserman, A. and Nellåker, C. (2018). Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. In: *European Conference on Computer Vision Workshops*, pp. 8–14. Springer.

Angwin, J. Larson, J. Mattu, S. and Kirchner, L. (2016). Machine Bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Aug, 2022.

Balkin, J. M. and Siegel, R. B. (Oct, 2003). The American Civil Rights Tradition: Anticlassification or Antisubordination. *Faculty Scholarship Series*, vol. 58, pp. 9–33.

Baltrušaitis, T. Robinson, P. and Morency, L.-P. (2016). Openface: An Open Source Facial Behavior Analysis Toolkit. In: *Winter Conference on Applications of Computer Vision*, pp. 1–10. IEEE.

Barocas, S. Hardt, M. and Narayanan, A. (Dec, 2017). Fairness in Machine Learning. *Neural Information Processing Systems Tutorial*, vol. 1, pp. 2.

Barocas, S. and Selbst, A. D. (Sep, 2016). Big Data's Disparate Impact. *California Law Review*, vol. 104, pp. 671.

Becker, D. V. Kenrick, D. T. Neuberg, S. L. Blackwell, K. and Smith, D. M. (Feb, 2007). The Confounded Nature of Angry Men and Happy Women. *Journal of Personality and Social Psychology*, vol. 92, pp. 179.

Berk, R. Heidari, H. Jabbari, S. Kearns, M. and Roth, A. (Jul, 2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, vol. 50, pp. 3–44.

Bettadapura, V. (2012). Face Expression Recognition and Analysis: The State of the Art. Retrieved from https://arxiv.org/abs/1203.6722. Aug, 2022.

Beutel, A. Chen, J. Doshi, T. Qian, H. Woodruff, A. Luu, C. Kreitmann, P. Bischof, J. and Chi, E. H. (2019). Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In: *Conference on AI, Ethics, and Society*, pp. 453–459. ACM.

Bhaskaruni, D. Hu, H. and Lan, C. (2019). Improving Prediction Fairness via Model Ensemble. In: *International Conference on Tools with Artificial Intelligence*, pp. 1810–1814. IEEE.

Biega, A. J. Gummadi, K. P. and Weikum, G. (2018). Equity of Attention: Amortizing Individual Fairness in Rankings. In: *Conference on Research Development in Information Retrieval*, pp. 1–10. ACM.

Böhlen, M. Chandola, V. and Salunkhe, A. (Nov, 2017). Server, Server in the Cloud. Who is the Fairest in the Crowd? Retrieved from https://arxiv.org/abs/1711.08801. Aug, 2022.

Bolukbasi, T. Chang, K.-W. Zou, J. Y. Saligrama, V. and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: *Advances in Neural Information Processing Systems*, pp. 1–9. Curran Associates, Inc.

Brown, G. Wyatt, J. Harris, R. and Yao, X. (Mar, 2005). Diversity Creation Methods: A Survey and Categorisation. *Information Fusion*, vol. 6, pp. 5–20.

Buhrmester, V. Münch, D. and Arens, M. (Oct, 2021). Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *Machine Learning and Knowledge Extraction*, vol. 3, pp. 966–989.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR.

Calders, T. and Verwer, S. (Jul, 2010). Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery*, vol. 21, pp. 277–292.

Caliskan, A. Bryson, J. J. and Narayanan, A. (Apr, 2017). Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, vol. 356, pp. 183–186.

Cash, T. F. and Kilcullen, R. N. (Jun, 1985). The Aye of the Beholder: Susceptibility to Sexism and Beautyism in the Evaluation of Managerial Applicants. *Journal of Applied Social Psychology*, vol. 15, pp. 591–605.

Castelnovo, A. Crupi, R. Greco, G. Regoli, D. Penco, I. G. and Cosentini, A. C. (Mar, 2022). A Clarification of the Nuances in the Fairness Metrics Landscape. *Scientific Reports*, vol. 12, pp. 1–21.

Caton, S. and Haas, C. (2020). Fairness in Machine Learning: A Survey. Retrieved from https://arxiv.org/abs/2010.04053. Aug, 2022.

Celis, L. E. and Keswani, V. (2019). Improved Adversarial Learning for Fair Classification. Retrieved from https://arxiv.org/abs/1901.10443. Aug, 2022.

Chakraborty, S. Tomsett, R. Raghavendra, R. Harborne, D. Alzantot, M. Cerutti, F. Srivastava, M. Preece, A. Julier, S. Rao, R. M. Kelley, T. D. Braines, D. Sensoy, M. Willis, C. J. and Gurram, P. (2017). Interpretability of Deep Learning Models: A Survey of Results. In: *SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, pp. 1–6. IEEE.

Chardon, A. Cretois, I. and Hourseau, C. (Aug, 1991). Skin Colour Typology and Suntanning Pathways. *International Journal of Cosmetic Science*, vol. 13, pp. 191–208.

Chen, Y. and Joo, J. (2021). Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In: *International Conference on Computer Vision*, pp. 14980–14991. IEEE.

Choi, Y. Choi, M. Kim, M. Ha, J.-W. Kim, S. and Choo, J. (2018). Stargan: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In: *Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797. IEEE.

Chouldechova, A. (Jun, 2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, vol. 5, pp. 153–163.

Corbett-Davies, S. and Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. Retrieved from https://arxiv.org/abs/1808.00023. Aug, 2022.

Corbett-Davies, S. Pierson, E. Feller, A. Goel, S. and Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. In: *International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM.

Crenshaw, K. (Jan, 1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, vol. 1989, pp. 139–167.

Cunningham, M. R. Roberts, A. R. Barbee, A. P. Druen, P. B. and Wu, C.-H. (Feb, 1995). “Their Ideas of Beauty are, on the Whole, the Same as Ours”: Consistency and Variability in the Cross-cultural Perception of Female Physical Attractiveness. *Journal of Personality and Social Psychology*, vol. 68, pp. 261.

Cutler, V. J. (Apr, 2021). The Science and Psychology of Beauty. *Essential Psychiatry for the Aesthetic Practitioner*, pp. 22–33.

Darwin, C. (2014). The Expression of the Emotions in Man and Animals. In: *Cambridge Library Collection*, pp. 398. University of Chicago Press.

Das, D. Sahoo, L. and Datta, S. (Feb, 2017). A Survey on Recommendation System. *International Journal of Computer Applications*, vol. 160, pp. 1–7.

Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Aug, 2022.

Datta, A. Tschantz, M. C. and Datta, A. (Mar, 2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Privacy Enhancing Technologies*, vol. 2015, pp. 92–112.

Denton, E. Hutchinson, B. Mitchell, M. Gebru, T. and Zaldivar, A. (2019). Image Counterfactual Sensitivity Analysis for Detecting Unintended Bias. In: *Conference on Computer Vision and Pattern Recognition Workshop on Fairness Accountability Transparency and Ethics in Computer Vision*, pp. 1–12. IEEE.

Drozdowski, P. Rathgeb, C. Dantcheva, A. Damer, N. and Busch, C. (Feb, 2021). Demographic Bias in Biometrics: A Survey on an Emerging Challenge. *IEEE Transactions on Technology and Society*, vol. 1, pp. 89–103.

Du, M. Yang, F. Zou, N. and Hu, X. (Jun, 2020). Fairness in Deep Learning: A Computational Perspective. *IEEE Intelligent Systems*, vol. 36, pp. 25–34.

Dunkelau, J. and Leuschel, M. (Oct, 2019). Fairness-Aware Machine Learning. *An Extensive Overview*, pp. 1–60.

Dwork, C. Hardt, M. Pitassi, T. Reingold, O. and Zemel, R. (2012). Fairness Through Awareness. In: *Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM.

Ekman, P. (1976). Pictures of Facial Affect. *Consulting Psychologists Press*.

Ekman, P. (1993). Facial Expression and Emotion. *American Psychologist*, vol. 48, pp. 384.

Ekman, P. and Friesen, W. (Sep, 1978). Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press, Palo Alto*.

Farkas, L. G. Hreczko, T. A. Kolar, J. C. and Munro, I. R. (Mar, 1985). Vertical and Horizontal Proportions of the Face in Young Adult North American Caucasians: Revision of Neoclassical Canons. *Plastic and Reconstructive Surgery*, vol. 75, pp. 328–338.

Fasel, B. (2002). Head-Pose Invariant Facial Expression Recognition Using Convolutional Neural Networks. In: *International Conference on Multimodal Interfaces*, pp. 529–534. IEEE.

Feffer, M. Hirzel, M. Hoffman, S. C. Kate, K. Ram, P. and Shinnar, A. (2022). An Empirical Study of Modular Bias Mitigators and Ensembles. Retrieved from https://arxiv.org/abs/2202.00751. Aug, 2022.

Fitzpatrick, T. B. (1975). Soleil et Peau. *J Med Esthet*, vol. 2, pp. 33–34.

Friedler, S. A. Scheidegger, C. and Venkatasubramanian, S. (Mar, 2021). The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM*, vol. 64, pp. 136–143.

Gan, J. Li, L. Zhai, Y. and Liu, Y. (Nov, 2014). Deep Self-Taught Learning for Facial Beauty Prediction. *Neurocomputing*, vol. 144, pp. 295–303.

Garg, P. Villasenor, J. and Foggo, V. (2020). Fairness Metrics: A Comparative Analysis. In: *International Conference on Big Data*, pp. 3662–3666. IEEE.

Ghani, N. A. Hamid, S. Hashem, I. A. T. and Ahmed, E. (Dec, 2019). Social Media Big Data Analytics: A Survey. *Computers in Human Behavior*, vol. 101, pp. 417–428.

Goodfellow, I. Bengio, Y. Courville, A. and Bengio, Y. (2016). *Deep Learning*. MIT press Cambridge.

Goodfellow, I. Pouget-Abadie, J. Mirza, M. Xu, B. Warde-Farley, D. Ozair, S. Courville, A. and Bengio, Y. (Dec, 2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680.

Goren, C. C. Sarty, M. and Wu, P. Y. (Oct, 1975). Visual Following and Pattern Discrimination of Face-like Stimuli by Newborn Infants. *Pediatrics*, vol. 56, pp. 544–549.

Gray, D. Yu, K. Xu, W. and Gong, Y. (2010). Predicting Facial Beauty Without Landmarks. In: *European Conference on Computer Vision*, pp. 434–447. Springer.

Grgić-Hlača, N. Zafar, M. B. Gummadi, K. P. and Weller, A. (2017). On fairness, Diversity and Randomness in Algorithmic Decision Making. In: *Fairness, Accountability, and Transparency in Machine Learning*, pp. 1–7. ACM.

Gunes, H. (2011). A Survey of Perception and Computation of Human Beauty. In: *Workshop on Human Gesture and Behavior Understanding*, pp. 19–24. ACM.

Haas, C. (2019). The Price of Fairness: A framework to Explore Trade-Offs in Algorithmic Fairness. In: *International Conference on Information Systems*. AIS.

Hardt, M. Price, E. and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In: *Advances in Neural Information Processing Systems*, pp. 3315–3323. MIT Press.

Hashimoto, T. Srivastava, M. Namkoong, H. and Liang, P. (2018). Fairness Without Demographics in Repeated Loss Minimization. In: *International Conference on Machine Learning*, pp. 1929–1938. PMLR.

He, K. Zhang, X. Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. In: *Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE.

Heusel, M. Ramsauer, H. Unterthiner, T. Nessler, B. and Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: *International Conference on Neural Information Processing Systems*, pp. 6629–6640. Curran Associates Inc.

Hönn, M. and Göz, G. (Jan, 2007). The Ideal of Facial Beauty: A Review. *Journal of Orofacial Orthopedics/Fortschritte der Kieferorthopädie*, vol. 68, pp. 6–16.

Hoofnagle, C. J. van der Sloot, B. and Borgesius, F. Z. (Fev, 2019). The European Union General Data Protection Regulation: What it is and What it Means. *Information & Communications Technology Law*, vol. 28, pp. 65–98.

Iancu, D. A. and Trichakis, N. (Jan, 2014). Pareto Efficiency in Robust Optimization. *Management Science*, vol. 60, pp. 130–147.

Jagielski, M. Kearns, M. Mao, J. Oprea, A. Roth, A. Sharifi-Malvajerdi, S. and Ullman, J. (2019). Differentially Private Fair Learning. In: *International Conference on Machine Learning*, pp. 3000–3008. PMLR.

Jain, A. Patel, H. Nagalapatti, L. Gupta, N. Mehta, S. Guttula, S. Mujumdar, S. Afzal, S. Sharma Mittal, R. and Munigala, V. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. In: *International Conference on Knowledge Discovery & Data Mining*, pp. 3561–3562. ACM.

Johnson, M. H. Dziurawiec, S. Ellis, H. and Morton, J. (Aug, 1991). Newborns' Preferential Tracking of Face-like Stimuli and its Subsequent Decline. *Cognition*, vol. 40, pp. 1–19.

Kagian, A. Dror, G. Leyvand, T. Cohen-Or, D. and Ruppin, E. (2007). A Humanlike Predictor of Facial Attractiveness. In: *Advances in Neural Information Processing Systems*, pp. 649–656. MIT Press.

Kamiran, F. and Calders, T. (Oct, 2012). Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems*, vol. 33, pp. 1–33.

Kamishima, T. Akaho, S. Asoh, H. and Sakuma, J. (2012). Fairness-aware Classifier with Prejudice Remover Regularizer. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer.

Kanade, T. Cohn, J. F. and Tian, Y. (2000). Comprehensive Database for Facial Expression Analysis. In: *International Conference on Automatic Face and Gesture Recognition*, pp. 46–53. IEEE.

Karkkainen, K. and Joo, J. (2021). Fairface: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In: *Winter Conference on Applications of Computer Vision*, pp. 1548–1558. IEEE.

Kenfack, P. J. Khan, A. M. Kazmi, S. A. Hussain, R. Oracevic, A. and Khattak, A. M. (2021). Impact of Model Ensemble On the Fairness of Classifiers in Machine Learning. In: *International Conference on Applied Artificial Intelligence*, pp. 1–6. IEEE.

Khan, A. Sohail, A. Zahoora, U. and Qureshi, A. S. (Apr, 2020). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artificial Intelligence Review*, vol. 53, pp. 5455–5516.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*, pp. 1–15. IEEE.

Klare, B. F. Burge, M. J. Klontz, J. C. Bruegge, R. W. V. and Jain, A. K. (Dec, 2012). Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 1789–1801.

Kleinberg, J. Ludwig, J. Mullainathan, S. and Rambachan, A. (2018). Algorithmic Fairness. In: *American Economic Association Papers and Proceedings*, vol. 108, pp. 22–27. American Economic Association.

Kleinberg, J. Mullainathan, S. and Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. Retrieved from https://arxiv.org/abs/1609.05807. Aug, 2022.

Kowner, R. (Jun, 1996). Facial Asymmetry and Attractiveness Judgement in Developmental Perspective. *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, pp. 662.

Langlois, J. H. Kalakanis, L. Rubenstein, A. J. Larson, A. Hallam, M. and Smoot, M. (May, 2000). Maxims or Myths of Beauty? A Meta-analytic and Theoretical Review. *Psychological Bulletin*, vol. 126, pp. 390.

Langlois, J. H. Roggman, L. A. Casey, R. J. Ritter, J. M. Rieser-Danner, L. A. and Jenkins, V. Y. (May, 1987). Infant Preferences for Attractive Faces: Rudiments of a Stereotype? *Developmental psychology*, vol. 23, pp. 363.

Li, S. and Deng, W. (2018). Deep Emotion Transfer Network for Cross-Database Facial Expression Recognition. In: *International Conference on Pattern Recognition*, pp. 3092–3099. IEEE.

Li, S. and Deng, W. (Jan, 2019). Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Transactions on Image Processing*, vol. 28, pp. 356–370.

Li, S. and Deng, W. (Mar, 2020). Deep Facial Expression Recognition: A Survey. *IEEE transactions on affective computing*, pp. 1–1.

Li, S. Deng, W. and Du, J. (2017). Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In: *Conference on Computer Vision and Pattern Recognition*, pp. 2584–2593. IEEE.

Lipton, Z. McAuley, J. and Chouldechova, A. (2018). Does Mitigating ML's Impact Disparity Require Treatment Disparity? In: *Advances in Neural Information Processing Systems*, pp. 8125–8135. MIT Press.

Little, A. C. (Sep, 2014). Facial Attractiveness. *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, pp. 621–634.

Liu, Z. Luo, P. Wang, X. and Tang, X. (2015). Deep Learning Face Attributes in the Wild. In: *International Conference on Computer Vision*, pp. 3730–3738. IEEE.

Loussaief, S. and Abdelkrim, A. (2016). Machine Learning Framework for Image Classification. In: *International Conference on Sciences of Electronics, Technologies of Information and Telecommunications*, pp. 58–61. IEEE.

Luong, B. T. Ruggieri, S. and Turini, F. (2011). K-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In: *International Conference on Knowledge Discovery and Data Mining*, pp. 502–510. ACM.

Ma, D. S. Correll, J. and Wittenbrink, B. (Dec, 2015). The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods*, vol. 47, pp. 1122–1135.

Ma, F. Sun, B. and Li, S. (Mar, 2021). Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing*, pp. 1–13.

Mavadati, S. M. Mahoor, M. H. Bartlett, K. Trinh, P. and Cohn, J. F. (Apr, 2013). Disfa: A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing*, vol. 4, pp. 151–160.

Mehrabi, N. Morstatter, F. Saxena, N. Lerman, K. and Galstyan, A. (Jul, 2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, vol. 54, pp. 1–35.

Mehrabian, A. (2017). Communication Without Words. In: *Communication Theory*, vol. 15, pp. 193–200. Routledge, 2 ed..

Mehrabian, A. and Russell, J. A. (1974). *An Approach to Environmental Psychology*. MIT Press.

Mollahosseini, A. Hasani, B. and Mahoor, M. H. (Jan, 2017). Affectnet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, vol. 10, pp. 18–31.

Moore, G. Galway, L. and Donnelly, M. (2017). Remember to Smile: Design of a Mobile Affective Technology to Help Promote Individual Happiness Through Smiling. In: *International Conference on Pervasive Computing Technologies for Healthcare*, pp. 348–354. ACM.

Nielsen, I. E. Dera, D. Rasool, G. Ramachandran, R. P. and Bouaynaya, N. C. (Jun, 2022). Robust Explainability: A Tutorial On Gradient-Based Attribution Methods for Deep Neural Networks. *IEEE Signal Processing Magazine*, vol. 39, pp. 73–84.

Osto, M. Hamzavi, I. H. Lim, H. W. and Kohli, I. (Jan, 2022). Individual Typology Angle and Fitzpatrick Skin Phototypes are Not Equivalent in Photodermatology. *Photochemistry and Photobiology*, vol. 98, pp. 127–129.

Peres, V. M. X. and Musse, S. R. (2021). Towards the Creation of Spontaneous Datasets Based on Youtube Reaction Videos. In: *International Symposium on Visual Computing*, pp. 203–215. Springer.

Perrett, D. I. Burt, D. M. Penton-Voak, I. S. Lee, K. J. Rowland, D. A. and Edwards, R. (Sep, 1999). Symmetry and Human Facial Attractiveness. *Evolution and Human Behavior*, vol. 20, pp. 295–307.

Pessach, D. and Shmueli, E. (Feb, 2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, vol. 55, pp. 1–44.

Pouyanfar, S. Yang, Y. Chen, S.-C. Shyu, M.-L. and Iyengar, S. S. (Jan, 2019). Multimedia Big Data Analytics: A Survey. *ACM Computing Surveys*, vol. 51, pp. 1–34.

Quadrianto, N. Sharmanska, V. and Thomas, O. (2019). Discovering Fair Representations in the Data Domain. In: *Conference on Computer Vision and Pattern Recognition*, pp. 8219–8228. IEEE.

Ramaswamy, V. V. Kim, S. S. and Russakovsky, O. (2021). Fair Attribute Classification Through Latent Space De-biasing. In: *Conference on Computer Vision and Pattern Recognition*, pp. 9301–9310. IEEE.

Revina, I. M. and Emmanuel, W. S. (Jul, 2021). A Survey on Human Face Expression Recognition Techniques. *Journal of King Saud University-Computer and Information Sciences*, vol. 33, pp. 619–628.

Rhodes, G. (Jan, 2006). The Evolutionary Psychology of Facial Beauty. *Annual Review of Psychology*, vol. 57, pp. 199–226.

Rhue, L. (Nov, 2018). Racial Influence on Automated Perceptions of Emotions. *Social Science Research Network*, pp. 1–11.

Rudin, C. (May, 2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, vol. 1, pp. 206–215.

Russakovsky, O. Deng, J. Su, H. Krause, J. Satheesh, S. Ma, S. Huang, Z. Karpathy, A. Khosla, A. Bernstein, M. et al. (Apr, 2015). Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, vol. 115, pp. 211–252.

Samuels, C. A. Butterworth, G. Roberts, T. Graupner, L. and Hole, G. (Oct, 1994). Facial Aesthetics: Babies Prefer Attractiveness to Symmetry. *Perception*, vol. 23, pp. 823–831.

Sattigeri, P. Hoffman, S. C. Chenthamarakshan, V. and Varshney, K. R. (Jul, 2019). Fairness GAN: Generating Datasets with Fairness Properties Using A Generative Adversarial Network. *IBM Journal of Research and Development*, vol. 63, pp. 3–1.

Schmid, K. Marx, D. and Samal, A. (Aug, 2008). Computation of a Face Attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition*, vol. 41, pp. 2710–2717.

Shen, W. and Liu, R. (2017). Learning Residual Images for Face Attribute Manipulation. In: *Conference on Computer Vision and Pattern Recognition*, pp. 4030–4038. IEEE.

Sherlock, M. and Wagstaff, D. L. (Oct, 2019). Exploring the Relationship Between Frequency of Instagram Use, Exposure to Idealized Images, and Psychological Well-being in Women. *Psychology of Popular Media Culture*, vol. 8, pp. 482–490.

Sigall, H. and Ostrove, N. (Mar, 1975). Beautiful But Dangerous: Effects of Offender Attractiveness and Nature of the Crime on Juridic Judgment. *Journal of Personality and Social Psychology*, vol. 31, pp. 410.

Simonyan, K. Vedaldi, A. and Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: *International Conference on Learning Representations*. IEEE.

Small, M. L. and Pager, D. (May, 2020). Sociological Perspectives on Racial Discrimination. *Journal of Economic Perspectives*, vol. 34, pp. 49–67.

Steephen, J. E. Mehta, S. R. and Bapi, R. S. (Feb, 2018). Do We Expect Women to Look Happier Than They Are? A Test of Gender-Dependent Perceptual Correction. *Perception*, vol. 47, pp. 232–235.

Thwaites, D. Lowe, B. Monkhouse, L. L. and Barnes, B. R. (Aug, 2012). The Impact of Negative Publicity on Celebrity Ad Endorsements. *Psychology & Marketing*, vol. 29, pp. 663–673.

Tian, Y. Kanade, T. and Cohn, J. F. (2011). Facial Expression Recognition. In: *Handbook of Face Recognition*, vol. 23, pp. 487–519. Springer, 2 ed..

Tian, Y.-I. Kanade, T. and Cohn, J. F. (Feb, 2001). Recognizing Action Units for Facial Expression Analysis. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, pp. 97–115.

Tomsett, R. Harborne, D. Chakraborty, S. Gurram, P. and Preece, A. (2020). Sanity Checks for Saliency Metrics. In: *AAAI Conference on Artificial Intelligence*, pp. 6021–6029. ACM.

Velusamy, S. Kannan, H. Anand, B. Sharma, A. and Navathe, B. (2011). A Method to Infer Emotions from Facial Action Units. In: *International Conference on Acoustics, Speech and Signal Processing*, pp. 2028–2031. IEEE.

Verma, S. and Rubin, J. (2018). Fairness Definitions Explained. In: *International Workshop on Software Fairness*, pp. 1–7. ACM.

Verrastro, V. Liga, F. Cuzzocrea, F. Gugliandolo, M. C. et al. (May, 2020). Fear the Instagram: Beauty Stereotypes, Body Image and Instagram Use in a Sample of Male and Female Adolescents. *Qwerty-Open and Interdisciplinary Journal of Technology, Culture and Education*, vol. 15, pp. 31–49.

Wang, T. Zhao, J. Yatskar, M. Chang, K.-W. and Ordonez, V. (2019). Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In: *International Conference on Computer Vision*, pp. 5310–5319. IEEE.

Wang, Z. Qinami, K. Karakozis, I. C. Genova, K. Nair, P. Hata, K. and Russakovsky, O. (2020). Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In: *Conference on Computer Vision and Pattern Recognition*, pp. 8919–8928. IEEE.

Xu, T. White, J. Kalkan, S. and Gunes, H. (2020). Investigating Bias and Fairness in Facial Expression Recognition. In: *European Conference on Computer Vision*, pp. 506–523. Springer.

Zhang, B. H. Lemoine, B. and Mitchell, M. (2018a). Mitigating Unwanted Biases with Adversarial Learning. In: *Conference on AI, Ethics, and Society*, pp. 335–340. ACM.

Zhang, D. Chen, F. Xu, Y. et al. (2016). *Computer Models for Facial Beauty Analysis*. Springer.

Zhang, Z. Luo, P. Loy, C.-C. and Tang, X. (2015). Learning Social Relation Traits From Face Images. In: *International Conference on Computer Vision*, pp. 3631–3639. IEEE.

Zhang, Z. Luo, P. Loy, C. C. and Tang, X. (May, 2018b). From Facial Expression Recognition to Interpersonal Relation Prediction. *International Journal of Computer Vision*, vol. 126, pp. 550–569.

Zhao, J. Wang, T. Yatskar, M. Ordonez, V. and Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: *Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989. ACL.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br