POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

Towards a Gender Inclusive Neural Network for Automatic Gender Recognition

Supervisors

Prof.sa Tania CERQUITELLI

Dr. Bartolomeo VACCHETTI

Candidate

Matteo BERTA

APRIL 2024

Summary

Today's data-driven systems and official statistics often oversimplify the concept of gender, reducing it to binary data, which carries far-reaching implications for policy development and equitable access to services. This simplification tends to result in misclassification and discrimination against individuals who identify as gender-nonconforming.

The proposed research aims to develop new, more equitable approaches that can effectively circumvent discrimination based on gender identity. Within this research framework, the primary emphasis lies in addressing the problem of underrepresentation and, in some instances, the complete absence of gender-nonconforming individuals in data collection efforts.

The work presented herein represents an initial endeavor to design an equitable neural network capable of accurately identifying gender within a multiclass context, including individuals whose gender identity transcends the binary spectrum. To achieve this objective, a comprehensive comparative analysis was conducted on several fine-tuned neural network models. The aim was to acquire a profound understanding of the pivotal distinguishing features in gender identification classification and to depict the limitations of current methodologies through the application of explainable AI techniques.

Different data collection techniques were tested in order to built a profitable dataset in a field where a suitable date collection is missing. The goal has been partially achieved, but the problem of invisibility is reflected in the availability of available and labeled data.

The fields of use that have been envisioned are those of inclusive communication. The creation of societal maps, useful for gaining representation, and the creation of a fairness score for cluster of images were the two main ideas for a fair use of the technology proposed.

The initial findings indicate promise, showcasing the efficacy of a fine-tuned EfficientNet model in precisely categorizing images of individuals according to their self-reported gender. However, there exists skepticism regarding its applicability in real-world scenarios due to the limited amount of data currently available concerning non-binary individuals.

Acknowledgements

"Je souhaitais que toute vie humaine
fût une pure liberté transparente"
Simone de Beauvoir

Table of Contents

List of Tables

List of Figures

Acronyms

AGR
Automatic Gender Recognition

AI
Artificial Intelligence

HCI
Human-Computer Interaction

CNN
Convolutional Neural Networks

RNN
Residual Neural Network

SVM
Support Vector Machines

GAN
Generative Adversarial Network

GPU
Graphics Processing Unit

TPU
Tensor Processing Unit

XAI
Explainable Artificial Intelligence

Gender Equality, Neural Network, Automatic Gender Recognition, Non-Binary

Chapter 1

Introduction

In an era dominated by technological advancements, the widespread use of data has become an integral part of daily lives. In the pursuit of social progress and justice, the imperative to address gender inequality has become a central concern across various disciplines. Despite the increasing awareness on societal problems and inclusiveness today's data-driven systems and official statistics often oversimplify the concept of gender, reducing it to binary data, with far-reaching implications for policy development and equitable access to services.

Being aware of the problem this thesis aims to bring a contribution on the proposed topic from technical perspectives with the not insignificant goal of helping to spread awareness of the problem.

We are working to advance our research in this area to develop new, more equitable approaches that can avoid discrimination based on gender identity. Within this research framework, our primary focus is on mitigating the problem of underrepresentation and, in some cases, the complete absence of non-binary individuals in data collection.

The goal of this thesis is to contribute to the development of a more equitable and less discriminatory Automatic Gender Recognition system. Recognizing the limitations inherent in current technologies and in the technology itself, we embarked on a comprehensive comparative analysis of various Automatic Gender Recognition (AGR) models. Our focus was on inclusivity, and to achieve this, we curated a self-built dataset comprising images of males, females, and non-binary individuals. This deliberate diversification aimed to create a more representative evaluation framework for AGR models, acknowledging the rich tapestry of gender expressions within our society.

An integral aspect of our methodology involved the application of an explainability model to discern the key features influencing the classification decisions of each gender class. Through this analytical lens, we sought to unravel the inner workings of AGR models, unveiling the factors that significantly shape their outcomes.

Our findings offer a glimpse into a promising future where Artificial Intelligence not only embrace inclusivity but also prioritize fairness in their classifications. However, our study has brought to light a crucial caveat: the imperative need for an evolution in data collection practices. The biases and limitations observed in current AGR systems underscore the necessity of expanding and diversifying the datasets used in their training. This revelation serves as a clarion call for a more conscientious approach to data collection, one that recognizes and encapsulates the diverse array of gender expressions and identities. In this pursuit, we envision the foundations of a technologically advanced yet socially responsible Automatic Gender Recognition system.

Chapter 2 is a descent down the rabbit hole of social invisibility of minorities, with a focus on gender inclusivity and fairness. There is a brief introduction to Automatic Gender Recognition (AGR) systems and an analysis of the social impact of the technology on LGBTQIA+ community.

Chapter 3 is a technical analysis of computer vision systems, face recognition models and automatic gender recognition.

Chapter 4 is the description of the methodological approach pursued during the preparation of the dissertation, and then, in Chapter 5, we move on to the analysis of the results obtained, with a distinction between quantitative and qualitative results and a focus on the comparison of our results and the sociological analysis found in Chapter 2.

The final chapter is the conclusion, in which a brief recap and a brief discussion will conclude the dissertation.

Chapter 2

Social Context

2.1 General definitions

To facilitate an accurate interpretation of the thesis, it is essential to provide clear and general definitions that mitigate the risk of misunderstanding. The objective of these definitions is to establish a common understanding of key terms and concepts, ensuring that the thesis is interpreted in the intended way.

Gender identity refers to a person's internal understanding of their own gender, which they privately experience in their sense of self. The definition of gender identity can be a complex and highly personal matter, as it may be consistent with or different from the sex assigned at birth based on physical sex characteristics. Each person's gender identity is unique and lies on a broad spectrum.

The term *non-binary* serves as an inclusive umbrella category that encompasses identities outside of traditional male and female gender identities or those that fall on a spectrum in between. The term *cisgender* refers to individuals whose gender identity aligns with the sex assigned at birth based on physical sex characteristics and it could be used in contrast with the term non-binary.

Genderqueer is a term used to describe a gender identity that doesn't fit within the traditional categories of male or female, often signaling a rejection of the gender binary. Individuals who identify as genderqueer may feel their gender is a mix, fluid, or entirely different from conventional notions. Genderqueer identities are diverse but share dis-identification with rigid gender binaries and in some cases, a direct challenge to the social institutions that perpetuate binaries [1]

The *gender spectrum* is a conceptual framework that acknowledges and represents the diversity of gender identities beyond the traditional binary understanding of male and female. Instead of viewing gender as a strict dichotomy, the gender spectrum recognizes that gender identity is a complex and multifaceted continuum. It encompasses a range of identities and expressions, allowing for variations that

go beyond the conventional categories of masculinity and femininity. The gender spectrum acknowledges that individuals may identify as a mix of genders, fall outside the binary entirely, or experience their gender identity fluidly over time. This inclusive concept aims to reflect the richness and diversity of human experiences related to gender.

Gender expression holds significant importance in this work, being defined as the behaviours linked to an outward manifestation of culturally established methods of conveying masculinity and/or femininity, or the deliberate rejection of these stereotypes.

Misgendering refers to the act of using language or pronouns that do not accurately reflect or respect an individual's gender identity. This can involve referring to someone using terms or pronouns associated with a gender other than the one with which they identify. Misgendering can be unintentional due to ignorance or oversight, but it can also be deliberate, and it can have negative emotional and psychological effects on the person being misgendered. Respecting and using the correct pronouns for individuals is an important aspect of affirming their gender identity. The concept of preferred pronouns is therefore essential to understand. The preferred pronouns are the pronouns an individual prefers to be addressed with, such as he/him, she/her, they/them, etc. The use of they as a singular pronoun for individuals who identify outside the gender binary is important.

Gender dysphoria is a psychological term used to describe the distress or discomfort that may arise when an individual's gender identity differs from the sex assigned at birth. In the context of gender dysphoria, gender marker, for instance the designation on official documents that indicates the legal recognition of an individual's gender, may be a source of distress if it does not align with an individual's gender identity.

2.2 Introduction to the societal problem

2.2.1 Social invisibility and intersectionality

Social invisibility refers to the condition in which certain individuals or groups within a society are overlooked, disregarded, or marginalized to the extent that their presence, experiences, and contributions go unnoticed or are downplayed. This phenomenon often stems from societal biases, discrimination, and ingrained stereotypes that perpetuate a narrow understanding of who is considered visible and important. Those affected by social invisibility may include minority ethnicities, genders, sexual orientations, and individuals from marginalized socio-economic backgrounds. The concept highlights the significance of recognizing and addressing the systemic factors that contribute to the erasure of certain groups, as well as the

impact of such invisibility on the well-being of individuals and communities. Social invisibility can lead to feelings of isolation, exclusion, and a lack of representation, reinforcing existing power imbalances. Addressing social invisibility requires proactive efforts to challenge stereotypes, promote inclusivity, and amplify the voices and experiences of those who have been historically marginalized. This advocacy aims to create a more equitable and just society where all individuals are acknowledged, valued, and afforded equal opportunities.

In 1989, Kimberlé Crenshaw authored "Demarginalizing the Intersection of Race and Sex," [2] marking a significant turning point in discussions pertaining to minorities and discrimination. Understanding the concept of intersectionality is essential as it recognizes that individuals belong to multiple social groups simultaneously. This recognition underscores the interconnected nature of systems of oppression and discrimination. Consequently, individuals may encounter distinct forms of discrimination when their various identities intersect. His study was focused on the limitations of legal frameworks in addressing the intersecting issues of race and gender discrimination. She highlighted the experiences of Black women who face marginalization by being excluded from both anti-racist and feminist remedies. The crucial takeaway from Crenshaw's work is the need to comprehend and internalize the idea that addressing the experiences of individuals with intersecting marginalized identities is vital for a thorough and effective approach to achieving social justice.

As Bragg, Renold, Ringrose and Jackson [3] showed the views of young people on gender diversity is significantly different from the ones perceived in the society in general. Younger generations have expanded vocabularies of gender identity and expression, improving their critical reflexivity about their own positions, and demonstrating a commitment to gender equality, gender diversity and the rights of gender and sexual minorities. This progressive shift in attitudes toward gender diversity among younger generations underscores the imperative for inclusive technology that not only reflects their nuanced understanding but also actively promotes an equitable and diverse digital landscape for individuals of all gender identities and expressions.

2.2.2 Technology and invisibility

Along with societal perceptions, technologies must evolve. In the context of continuing progress assigning gender as a binary category in coding practices creates significant socio-technical challenges, as it inherently excludes individuals who identify outside the traditional binary framework (i.e., beyond the categories of male and female). This approach poses technical barriers, hindering the ability of non-binary and gender-diverse individuals to register for services and accurately participate in data collection exercises. Even though recent research suggests that approximately 1.6% of U.S. adults identify as non-binary [4], the entrenched nature

of gender binarism within institutionalized heterosexism poses significant challenges for those seeking to live outside conventional norms. This statistic highlights a growing recognition of diverse gender identities, yet the societal structures, deeply ingrained in binary systems, continue to create obstacles for individuals pursuing alternative ways of living and expressing their gender. The intersectionality of gender identity with institutionalized heterosexism underscores the need for broader societal shifts to foster inclusivity, understanding, and acceptance of the multiplicity of gender experiences [1]. For the reason expressed above a study conducted in UK in 2018 showed that 76% of non-binary people avoided expressing their gender identity due to fear of negative reactions [5]. By adhering strictly to a binary coding system, systems fail to accommodate the diverse and nuanced spectrum of gender identities. This oversight not only overlooks the lived experiences of non-binary, genderqueer, and other gender-nonconforming individuals but also perpetuates a digital environment that marginalizes and erases their existence.

Joni Seager said, "what gets counted count". With this simple, yet aggressive and meaningful sentence the geographer pointed out the enormous issue of being excluded by official datasets. Excluding a whole category of individuals from official datasets has broad and profound impacts on societal dynamics. This exclusion leads to the underrepresentation of marginalized groups, perpetuating stereotypes and biases while hindering accurate data analysis. Policymakers may lack crucial insights, resulting in inequitable resource allocation and flawed decision-making. The absence of a category raises ethical and civil rights concerns, contributing to the systemic marginalization of certain communities. The exclusion of a category from official datasets extends its impact beyond social repercussions, reaching into realms like innovation and public health. By omitting certain groups, innovation is stifled as diverse perspectives and ideas are overlooked. Advocating for inclusive data collection practices is imperative to cultivate a more equitable society. Inclusivity, beyond fostering social justice, ensures that policies, resources, and technologies respond effectively to the diverse needs and experiences of all individuals. This advocacy is a crucial step towards building a more just and inclusive society, where the richness of perspectives is acknowledged and utilized for advancements in various fields, from technology to public health. In essence, embracing inclusivity in data collection becomes a cornerstone for progress, enabling a society that truly caters to the diverse realities of its members.

People outside the gender binary are a small percentage of the population, and under this excuse they are almost always excluded from the data collection, but this is a logic mistake, as the trans activist Maria Munir highlighted: "If you refuse to register non-binary people like me with birth certificates, and exclude us in everything from creating bank accounts to signing up for mailing lists, you do not have the right to turn around and say that there are not enough of us to warrant change" [6]. Our work as data scientists should focus on the development

of better practices to deal with outliers and minority population, to ensure a fair and less discriminatory representation of the world through datasets, respecting self-identification and privacy [6].

As Jennifer Rode [7] found exploring HCI's models of gender they "assume gender is stagnant and based on physiology", highlighting the need to "move past binary gender in order to allow flexible discussion of gender and technology".

2.2.3 Mental Health for Non-Binary Individuals

It is important to notice that non-binary individuals may experience greater risk for negative mental health than cisgender heterosexual people [8]. In addition to the stress created by discrimination they might experience added stress due to others consistently using incorrect gender pronouns to refer to them and the frequent need to disclose their non-binary identity [9]. They also encounter elevated levels of discrimination stemming from societal repercussions related to their refusal to conform to the gender binary. Moreover, they face both macro and micro aggressions, subtly suggesting their invisibility, questioning their legitimacy, or requiring them to continually justify their identity [10].

The experience of being misgender by people has been shown to prime individuals for rejection, impact self-esteem and felt authenticity, and increase one's perception of being socially stigmatised. Being misgendered by a machine could be, potentially, worse, as it can contribute to the perpetuation of gender stereotypes and reinforce biases that already exist in society [11]. This could potentially lead, in addition to psychological consequences for those involved, to a lack of trust in the technology and a rejection of it. An interesting study on the perception of genderqueer people on AGR systems [12] the participants involved answered that being misgendered by a machine is worse than being misgendered by a human saying, "I get misgender enough by human beings, why on Earth would I want a robot to help in that?" or "Not only is human error getting my identity false, it's computer and Ais and technology also messing up too. It's not a person's uncertain perception, it's a more precise mathematical analysis of me that led to this conclusion, which kinda would run it in my face even more".

Being aware of the psychological implications of what will be covered in this dissertation, the goal is to be able to help decrease the psychological pressure exerted by these types of algorithms, making them fairer and less discriminatory for the purpose of greater inclusion.

2.3 Automatic Gender Recognition systems

2.3.1 Gender data

In the sphere of social research and policy formulation, gender data assumes a crucial role. It regards the collection and analysis of information based on gender. This type of data provides interest information in revealing disparities between genders, guiding policymakers in tailoring targeted interventions for issues spanning education, healthcare, and employment.

Gender is considered a soft biometric data and in a related context, the significance of soft biometric data, encompassing features like gender expression, is on the rise, especially within fields such as facial recognition technology. Unlike traditional biometrics that focus solely on unique physical traits, soft biometrics, including gender data, contribute to a more holistic understanding of individuals without compromising personal identity. This nuanced approach proves valuable across diverse applications, from security systems to user interface customization, where the acknowledgment and respect of gender expression are increasingly deemed essential.

"Gender data" typically refers to information collected and analyzed in the context of gender, encompassing various aspects related to the experiences, roles, and identities of individuals. This data can include demographic information, such as the distribution of individuals across gender categories, as well as details about socio-economic factors, education, health, and participation in various fields based on gender.

Gender data is crucial for understanding and addressing disparities, inequalities, and discrimination based on gender. It helps policymakers, researchers, and organizations develop informed strategies and policies to promote gender equity and inclusivity. Additionally, gender data may involve information on gender-based violence, representation in leadership roles, and other factors that contribute to a comprehensive understanding of how gender operates in different societal contexts. The ability to accurately determine an individual's gender is also important for expanding advertising categories. The need of expanding advertising categories is clearly showed by the fact that Facebook has increased its gender options for English speakers in the UK and US from 2 to 58, but still groups all genders into a binary system within advertising categories, highlighting the challenge of visibility in this context. [13]. Automatic Gender Recognition Systems (AGR) based on facial images are technological solutions that leverage machine learning and computer vision techniques to identify and categorize an individual's gender. These systems rely on algorithms, often employing deep learning models like Convolutional Neural Networks, to analyse facial features such as face shape, jawline, and cheekbones. The training process involves using large datasets of labelled facial images, enabling

the algorithm to learn patterns that correlate with gender.

2.3.2 History of AGR

During years AGR systems have crossed several phases. In the early 2000s, facial recognition technology emerged, attempting to incorporate gender classification based on facial features. The 2010s witnessed a significant leap with the integration of deep learning techniques. This integration significantly improved accuracy of every computer vision system, including facial recognition and, consequently, gender prediction.

Despite their potential accuracy, AGR systems face challenges and limitations. Accuracy rates may vary among different demographic groups, and biases may be present due to imbalances in the training data. A major concern is the reinforcement of gender stereotypes, as the algorithms learn from societal norms embedded in the training data. Ethical issues arise, particularly in terms of privacy and consent, as deployment without informed consent infringes on privacy rights. Cultural variations in facial features also pose challenges, as AGR systems need to account for diversity across ethnicities and cultures. In addition, it is interesting to notice why AGR is necessary, because most AGR papers, according to (misgendering machines) do not dedicate any time to discussing the problem this technology is a solution to. To align with new legislations, such as the European General Data Protection Regulation (GDPR), and uphold ethical standards, there has been a notable deceleration in the utilization of this technology. Microsoft, for example, has retired facial recognition capabilities that can be used to try to infer emotional states and identity attributes from their Azure AI Face services [14]. In the document titled "Responsible AI Investments and Safeguards for Facial Recognition," issued on June 21, 2022, Microsoft articulated its commitment to ethical practices and responsible deployment of artificial intelligence. “We will retire facial analysis capabilities that purport to infer emotional states and identity attributes such as gender, age, smile, facial hair, hair, and makeup. We collaborated with internal and external researchers to understand the limitations and potential benefits of this technology and navigate the tradeoffs. In the case of emotion classification specifically, these efforts raised important questions about privacy, the lack of consensus on a definition of “emotions,” and the inability to generalize the linkage between facial expression and emotional state across use cases, regions, and demographics. API access to capabilities that predict sensitive attributes also opens up a wide range of ways they can be misused—including subjecting people to stereotyping, discrimination, or unfair denial of services [. . .] In line with the supporting goals and requirements outlined in the Responsible AI Standard, we are bolstering our investments in fairness and transparency. We are undertaking responsible data collections to identify and mitigate disparities in the performance

of the technology across demographic groups and assessing ways to present this information in a way that would be insightful and actionable for our customers." [15].

2.3.3 The Misgendering Machines

An interesting point of view is carried out by Os Keyes, a PhD candidate at the University of Washington's Department of Centred Design & Engineering, in his dissertation named "The Misgendering Machines". Highlighting the difficulties of Human-Computer Interactions to consider and include transgender and non-binary perspective in research, they focused on the implication of AGR for trans people focusing on the fact that AGR "consistently operationalises gender in a trans-exclusive way".

The analysis is conducted by a content analysis of different AGR literature and on Human-Computer Interaction papers that rely on AGR technology with the results that AGR research mostly ignores the existence of transgender people, with erasure and no discussion of the category, reflecting the erasure perpetuated by the society. In particular, the analysis showed that, analysing 58 AGR papers, 95% of them treat gender as binary, 72% as immutable and 60% as physiological. The 12 examined papers about Human-Computer Interactions do not mention non-binary individuals and give no indication that gender is more complex than the traditional binary model, directly or indirectly.

The work by Os Keyes clearly shows a problem within the research community about computer vision by showing a lack of awareness regarding the proposed topic or, in part, a willingness to ignore the complexity of the genre so as to limit the technical complexity inherent in developing a face recognition system.

2.4 Gender as Performative Act

2.4.1 Gender as an historical situation

"One is not born, but rather becomes, a woman" [16] is ones of the most famous de Beauvoir's statement and is an important distinction between sex and gender. In the "Second Sex", published in 1949, the French philosopher introduced the concept of gender as an historical situation rather than a natural fact. The idea is to challenge the notion that being a woman or a man is predetermined or natural. According to de Beauvoir, gender identities are socially and culturally constructed through historical processes and societal norms. The roles and expectations assigned to each gender are contingent on the specific historical context in which they emerge. This perspective implies that gender roles are not static, but they evolve over time in response to historical developments, cultural shifts, and societal changes.

Understanding the historical situation of gender allows us to recognize the fluidity of gender identities and roles. To comprehend the situation of an individual in a particular society, one must consider the historical context that has shaped and continues to shape their roles, rights and status. This historical situation encompasses legal, economic, political and cultural factors that contribute to the construction of gender norms.

The concept expressed above, extended to our contemporary times, must be interlaced with the concept of intersectionality introduced by Crenshaw [2], because different groups of individuals may experience their historical situation in distinct ways due to the interplay of these intersecting factors. This perspective on "gender as an historical situation" not only challenges essentialist [17] views of gender but also prompts a deeper exploration into the complex web of forces that contribute to the perpetuation of gender inequalities over time. By acknowledging that gender roles are not fixed and are intricately woven into the fabric of historical processes and societal norms, scholars and activists are urged to conduct a critical examination of the multifaceted elements that sustain gender disparities.

This critical examination involves scrutinizing not only explicit laws and policies but also the implicit, often subtle, mechanisms embedded in cultural practices, economic structures, and political frameworks. It encourages an exploration of historical developments that have shaped prevailing attitudes toward gender and the ways in which these attitudes have been institutionalized. By understanding the historical context of gender, we gain insights into how certain norms and expectations have been constructed and perpetuated, influencing the lived experiences of individuals. The emphasis on intersectionality adds another layer to this examination, acknowledging that the historical situation of gender cannot be isolated from other social categories like race, class, and ethnicity. Different groups within society may experience and negotiate their historical gender situations in unique ways, emphasizing the need for nuanced and inclusive analyses that consider the interplay of various factors.

In essence, the call for a critical examination of the forces shaping gender inequalities over time encourages a holistic understanding of the historical situation of gender. This concept, the incipit of those we will explain in a moment, is necessary to fully understand the motivations behind the work proposed in this dissertation.

2.4.2 Gender as an act of performance

In the late ‘80s West and Zimmerman proposed their study called “Doing Gender” [17] with the purpose of “propose and ethnomethodologically informed understanding of gender as a routine, methodical, and recurring accomplishment.” Their idea was a conception of gender as an emergent feature of social situations composed by

an outcome and a rationale for various social arrangements.

It is important to underline that the authors focused on a binary view of gender, but the concept of "doing gender" could be easily expanded to a complete view of gender. "Doing gender" means creating differences between genders, differences that are not biological or essential, and then use these differences to reinforce the "essentialness" of gender. The authors pointed out that "any social encounter can be pressed into service in the interests of doing gender", so a person's gender is something that one does, and does recurrently, in interaction with others.

The concept of gender as a performative act is rooted in the work of Judith Butler [18], a prominent philosopher and gender theorist. Butler's ideas have significantly influenced the field of gender studies, challenging traditional notions of gender as a fixed and inherent aspect of identity. Before Butler's work, discussions around gender often centered on the assumption that gender identity was a stable and natural essence linked to biological sex. However, Butler contested this idea, proposing that gender is not something one inherently possesses but rather something one continually performs through a series of repeated actions.

Butler's theory of gender performativity suggests that gender identity is not a pre-existing reality but rather a socially constructed and enacted performance. In other words, individuals engage in specific behaviors, gestures, and expressions that align with societal expectations of masculinity or femininity, and it is through these repetitive performances that gender identity is produced and maintained. Butler draws on the work of philosopher and linguist J.L. Austin, who introduced the concept of "performative utterances" [19], speech acts that not only describe, but also perform an action. Butler extends this idea to the realm of gender, arguing that gender identity is constituted through a series of performative acts, including language, gestures, and other bodily expressions.

Moreover, Butler's work has been instrumental in deconstructing the traditional heteronormative framework surrounding gender [1]. By highlighting the performative nature of gender, she provides a theoretical foundation for understanding and challenging the limitations imposed by societal expectations. The concept of performativity extends beyond individual actions to encompass broader social structures and institutions that perpetuate normative gender roles.

Butler's work lays the foundation for the idea of the work proposed in this dissertation and justifies the attempt to improve AGR systems.

Chapter 3

Computer Vision

This chapter offers an introduction to important concepts and advancements in deep learning, focusing on computer vision and automatic gender recognition systems. Reviewing the history of computer vision systems from their inception to the present day and dwelling on the most innovative models, the models used in our comparative analysis will be introduced so that the rest of the dissertation is understandable and clear.

A systematic review of relevant literature will be presented, with a special focus on residual networks and convolutional neural networks (CNNs), highlighting the innovations of the last decade in the field of computer vision.

After reading this chapter you will have the tools necessary to fully understand the methodology used to conduct the proposed study.

Computer vision is a dynamic and interdisciplinary field that strives to empower machines with the ability to interpret and comprehend visual information in a manner similar to human vision. It encompasses a broad spectrum of techniques, methodologies, and technologies aimed at endowing computers with the capacity to extract meaningful insights from images and videos, thereby enabling them to make informed decisions and perform tasks that traditionally rely on human visual perception.

At its core, computer vision involves the development and implementation of algorithms and systems that facilitate the extraction, analysis, and interpretation of features and patterns present in visual data. These algorithms leverage concepts from various domains, including computer science, mathematics, physics, and neuroscience, to enable machines to navigate the complexities of the visual world.

The fundamental objectives of computer vision include, but are not limited to:

1. **Image Recognition:** Teaching machines to identify and classify objects, scenes, or patterns within images.

2. **Object Detection:** Enabling machines to locate and delineate specific objects or entities within a given visual context.

3. **Facial Recognition:** Allowing machines to identify and verify individuals based on facial features.

4. **Scene Understanding:** Providing machines with the capability to comprehend the overall context and meaning of a scene depicted in visual data.

5. **Image Generation:** Empowering machines to create new visual content, such as generating realistic images or enhancing existing ones.

6. **Video Analysis:** Extending the principles of computer vision to the temporal domain, enabling machines to analyze and understand videos.

The field leverages advancements in technologies such as machine learning, particularly deep learning, where convolutional neural networks (CNNs) have proven particularly effective in handling complex visual tasks. The integration of artificial intelligence further enhances the adaptability and decision-making capabilities of computer vision systems.

The applications of computer vision are diverse and continue to expand across various industries. In healthcare, computer vision aids in medical image analysis and diagnosis. In autonomous vehicles, it plays a pivotal role in recognizing and understanding the surrounding environment. In security and surveillance, it facilitates real-time monitoring and threat detection. Moreover, computer vision contributes to the immersive experiences of virtual reality and augmented reality, transforming the way humans interact with digital environments.

3.1 History of Computer Vision

- **1950s-1960s: Inception and Early Exploration**
  The origins of computer vision date back to the 1950s and 1960s, marked by pioneering efforts to imbue computers with the ability to comprehend and interpret visual data. During this period, researchers delved into the conceptualization of teaching computers to understand the visual world. Frank Rosenblatt's creation of the "Perceptron," an early artificial neural network, laid the groundwork for subsequent advancements in neural network-based computer vision systems.

- **1970s-1980s: Foundational Algorithms for Image Analysis**
  In this era, the focus shifted towards the development of algorithms designed for edge detection and feature extraction from images. Fundamental to more sophisticated computer vision tasks, significant contributions included the

introduction of the "Canny edge detector" – an algorithm employing a multi-stage approach to identify a broad spectrum of edges in images. Additionally, the "Hough transform," a technique for detecting simple geometric shapes in images, marked a crucial development during this period.

- **1980s-1990s: Progress in Object Recognition and Machine Learning**
  In this era, the focus shifted towards the development of algorithms designed for edge detection and feature extraction from images. Fundamental to more sophisticated computer vision tasks, significant contributions included the introduction of the "Canny edge detector" – an algorithm employing a multi-stage approach to identify a broad spectrum of edges in images. Additionally, the "Hough transform," a technique for detecting simple geometric shapes in images, marked a crucial development during this period.

- **2000s: Rise of Support Vector Machines and Viola-Jones Algorithm**
  Support Vector Machines (SVMs) emerged as a popular choice for object recognition tasks during this period. The Viola-Jones object detection framework, utilizing AdaBoost, became widely adopted for real-time face detection.

- **2000s-2010s: Deep Learning Revolutionizes Computer Vision**
  The late 2000s witnessed a paradigm shift with the breakthrough in deep learning, notably the ascendancy of Convolutional Neural Networks (CNNs). CNNs demonstrated remarkable efficacy in image classification tasks, a transformation exemplified by influential architectures like AlexNet, VGGNet, and ResNet. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) played a pivotal role in propelling deep learning within the realm of computer vision, opening doors to applications such as object detection, image segmentation, and image generation.

- **2010s-Present: Ongoing Advancements and Diversification**
  Continuing into the present, computer vision has seen ongoing progress marked by the prevalence of transfer learning techniques, including the fine-tuning of pre-trained models. Generative Adversarial Networks (GANs) have been instrumental in generating realistic images and videos. Attention mechanisms, as epitomized by Transformer models, have found applications in various computer vision tasks. Furthermore, the development of specialized hardware such as Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) has significantly accelerated the training of deep neural networks for computer vision applications. [20]

3.2 Computer Vision Models

In this chapter, our attention is directed towards specific computer vision models featured in our comparative analysis. Emphasis will be placed on the rationale behind selecting these models and an exploration of the innovations inherent to each. The models under consideration include ResNet [21], MobileNet [22], Inception Models [23], and EfficientNet [24].

3.2.1 Resnet

ResNet, short for Residual Networks, represents a significant breakthrough in the field of deep learning and computer vision. Developed by researchers Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun at Microsoft Research in 2015 [21], ResNet introduced a novel architectural design that addressed the challenges associated with training very deep neural networks.

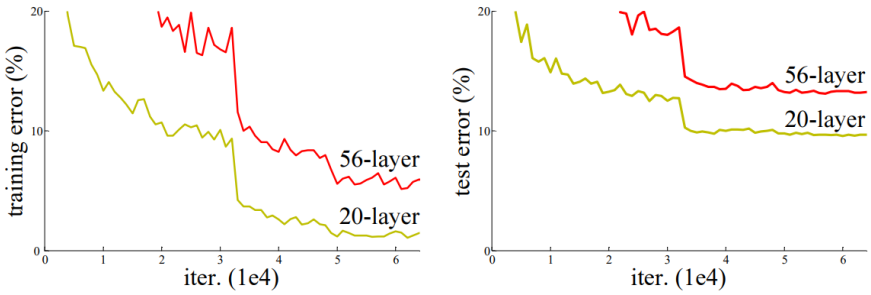


Figure 3.1: Training error and test error with 20 layers and 56 layers. Source [21]

One of the primary issues faced by deep neural networks is the degradation problem. As the depth of a neural network increases, the performance on the training set saturates and then degrades rapidly. As visible in 3.1 the deeper network analyzed in [21] has higher training and test error. This phenomenon contradicts the common intuition that a model's performance should improve as it becomes more complex. The degradation problem is primarily attributed to the difficulty of training deep networks and the vanishing gradient problem explained in 1994 by Bengio, Simard and Frasconi [25], where gradients diminish or explode as they are backpropagated through the layers during training.

ResNet addresses this problem by introducing residual learning. The key idea is to introduce shortcut connections, also known as skip connections or identity mappings, that allow the network to directly learn the residual mapping—the difference between the input and output of a given layer. This concept is based

on the observation that it is often easier for a neural network to learn a residual mapping than to learn the entire mapping from scratch.

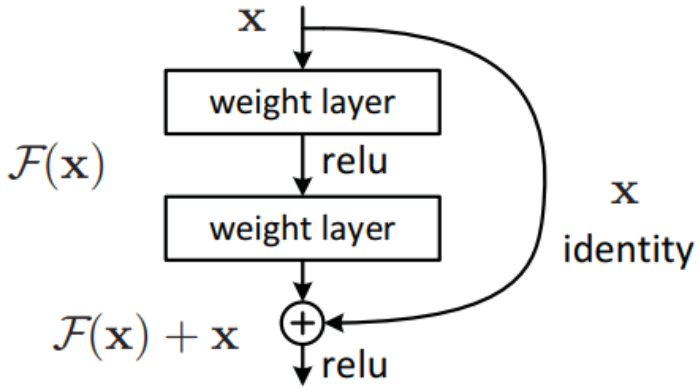


Figure 3.2: Building block of residual learning. Source [21].

The residual block, which is the building block of ResNet, consists of two main paths: the identity path, which simply passes the input to the output without any change, and the residual path, which learns the residual mapping. An example of residual block is shown in 3.2. Mathematically, the output of a residual block (H(x)) is given by:

$$H(x) = F(x) + x$$

Here, $F(x)$ represents the residual mapping learned by the layers within the block, and x is the input to the block. By introducing this skip connection, the gradient can flow directly through the identity path during backpropagation, mitigating the vanishing gradient problem and facilitating the training of very deep networks.

The architecture comprises a stack of residual blocks, and the overall structure is organized into multiple stages. Each stage typically consists of several residual blocks with a consistent number of filters, and the spatial resolution is reduced while the number of filters is increased. This design helps to maintain a balance between capturing high-level features and preserving spatial information.

The concept of identity mapping is sufficient for addressing the degradation problem and is economical and it is made by a shortcut connections:

$$y = F(x, W_i) + x$$

. In the equation expressed above x and y are input and output vectors of the layer considered, while the function

$$F(x, W_i)$$

represents the residual mapping to be learned.

ResNet's impact extends beyond its ability to train very deep networks. It has become a cornerstone in various computer vision tasks, including image classification, object detection, and image segmentation. Pre-trained ResNet models, especially those trained on large datasets like ImageNet, have been widely used for transfer learning, where a pre-trained model is fine-tuned on a specific task with a smaller dataset, as in the case proposed in this dissertation.

This idea offers the possibility to win the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2015 thanks to the ability of training deeper neural network addressing the vanishing gradient problem.

In conclusion, ResNet has had a profound impact on the field of deep learning by introducing a groundbreaking architecture that enables the training of very deep neural networks. The incorporation of residual learning and skip connections has proven effective in mitigating the degradation problem, allowing for improved performance on a wide range of computer vision tasks. ResNet's principles have influenced subsequent architectures and continue to shape the landscape of deep neural network design.

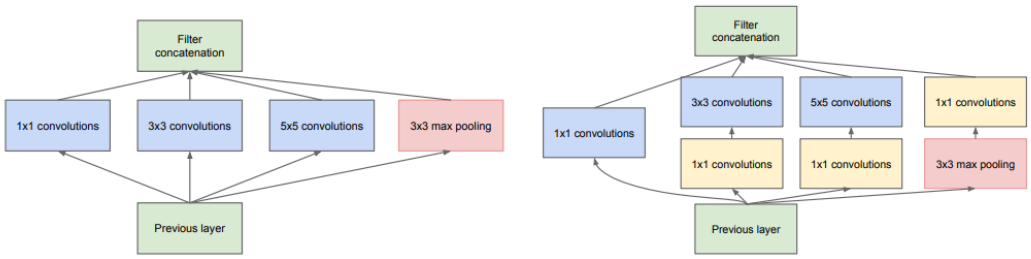
3.2.2 Inception models

Enhancing the performance of deep neural networks is most effectively achieved by enlarging their dimensions, both by increasing the depth, formally the number of levels of the network and the width, the number of units at each level.

Despite the possibility to train higher quality models with this solution there are some drawbacks. The first one is the larger number of parameters that could lead to overfitting especially if the number of labeled examples in the training set is limited, creating the problem of the need of high quality and expensive training sets. A second drawback is the increased use of computational resources. For example, if two convolutional layers are chained any uniform increase in the number of their filters results in a quadratic increase of computation. The theory tell that both the problems could be solved moving from fully connected to sparse architectures, but todays computer infrastructures are inefficient on numerical calculation on non-uniform sparse data structures.

The idea behind the Inception Models proposed in 2014 [23] is to create an "architecture that makes use of the extra sparsity, even at filter level, but exploits our current hardware by utilizing computations on dense matrices".

At the core of Inception models are the inception modules, shown in 3.3. These modules consist of parallel convolutional operations with different filter sizes, facilitating the capture of features at various scales within the same layer. The inception module essentially acts as a multi-path architecture, enabling the network to learn a diverse set of features in parallel. Key components of an inception module



(a) Inception module, naïve version (b) Inception module with dimension reductions

Figure 3.3: Inception module. Source [23].

include 1x1 convolutional filters, 3x3 convolutional filters, 5x5 convolutional filters, and max pooling, each contributing to the model's ability to understand complex patterns efficiently.

A noteworthy aspect contributing to the computational efficiency of Inception models is the use of 1x1 convolutions. These convolutions serve as bottleneck layers, reducing the number of input channels and computational complexity. By incorporating 1x1 convolutions strategically, the model can maintain depth while preserving spatial information, thus optimizing the use of computational resources. The idea is shown in figure 3.3(b) and could be summarized in the use of 1x1 convolutions to compute reductions before the expensive 3x3 and 5x5 convolutions.

The complete architecture proposed is shown in Figure 3.4 and it comprises 22 layers (27 if we also count pooling), with 9 inception modules.

An intriguing observation lies in the notable effectiveness of relatively shallow networks for this task. This implies that the features generated by middle layers of the network possess high discriminative power. Introducing auxiliary classifiers linked to these intermediate layers is anticipated to promote discrimination in the earlier stages of the classifier. This approach aims to amplify the gradient signal during backpropagation and contribute supplementary regularization to the model.

The success of GoogleNet marked a turning point in deep learning for image classification, inspiring subsequent iterations and improvements in the Inception architecture. Subsequent versions, including Inception V2, V3, and V4, introduced enhancements such as batch normalization, factorized convolutions, and integration of ideas from ResNet architectures for improved accuracy and efficiency.

In particular we focused on the use of Inception V3 [26]. Compared to GoogLeNet (Inception V1) [23] the new version includes improvements such as factorized convolutions, batch normalization, and an increased number of layers, making it deeper and more complex and introduced batch normalization, a technique that helps stabilize and accelerate the training process by normalizing inputs.

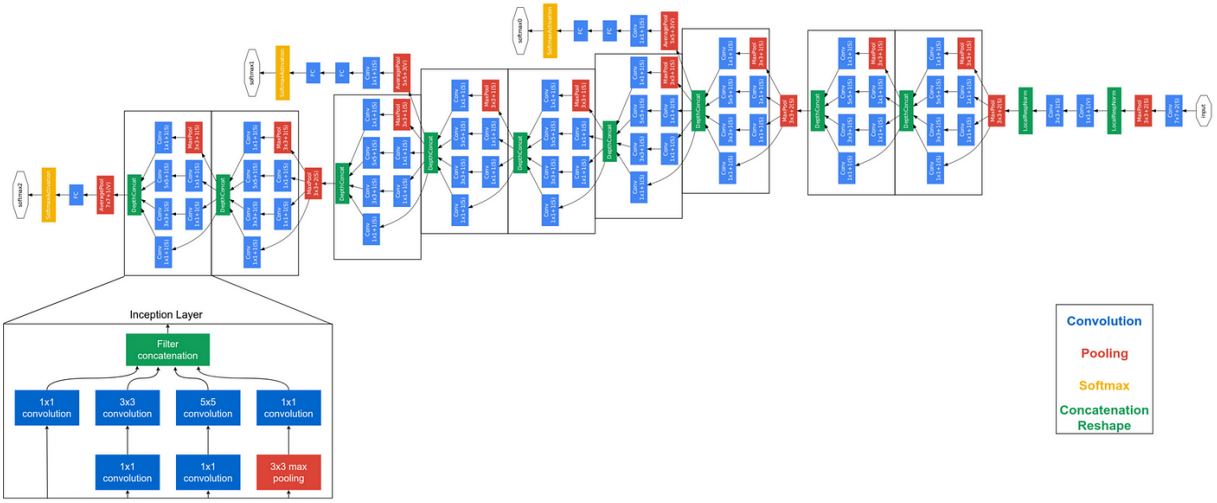


Figure 3.4: Complete structure of GoogLeNet network. Source [23].

3.2.3 MobileNets

MobileNet [22] stands as a revolutionary convolutional neural network architecture engineered explicitly for mobile and embedded vision applications. It represents a groundbreaking effort by Google researchers to enhance deep learning models for devices with limited computational resources, a crucial advancement in response to the growing need for on-device AI capabilities.

At the heart of MobileNet lies its innovative utilization of depthwise separable convolutions, a departure from conventional convolutional techniques. This transformative approach divides the convolutional process into two distinct phases: the depthwise convolution and the pointwise convolution. The former operates by applying a single convolutional filter per input channel, independently processing spatial information across channels. Subsequently, the pointwise convolution, executed as a 1x1 convolution, amalgamates the outputs from the depthwise convolution, fostering inter-channel information integration. Such modular convolutional operations, streamlined and efficient, form the cornerstone of MobileNet's architecture, distinguishing it from its predecessors. The representation of this concept is visibile in figure 3.5.

MobileNet uses 3x3 depthwise separable convolutions which uses between 8 to 9 less computation than standard convolutions at only a small reduction in accuracy, as visibile in 3.6. MobileNet introduced two parameter, the *width multiplier* α and the *resolution multiplier* ρ. The first one is studied to thin a network uniformly at each layer. For a given layer and given width multiplier α, the number of input channels M becomes αM and the number of output channels N becomes αN. Being α equal to 1 for the baseline MobileNets and less than 1 for the reduced MobileNets

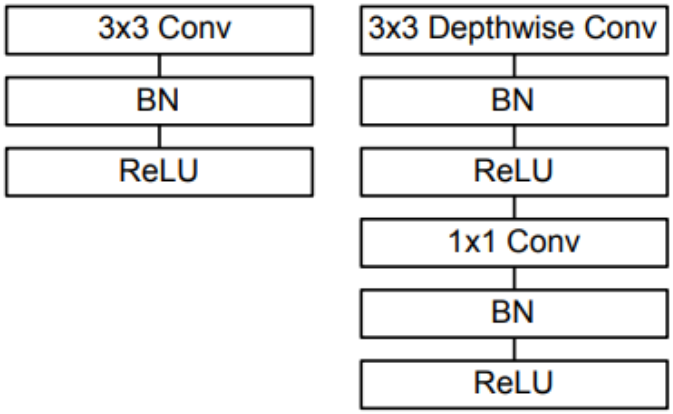


Figure 3.5: Left: Standard convolutional layer with batchnorm and ReLU. Right: Dephtwise Separable convolutions with Depthwise and Pointwise layers followed by batchnorm and ReLU. Source [22].

the computational cost is reduced significantly and used wisely to evaluate a tradeoff between computational cost and accuracy. The second hyperparameter to reduce the cost of a neural network is the resolution multiplier ρ. It is applied to the input image and the internal representation of every layer is reduce by the same multiplier, so it is set by setting the input resolution. The effect is to reduce the computational cost by ρ^2.

In conclusion MobileNet enables fast and accurate image classification and object detection tasks with reduced computational requirements compared to traditional convolutional neural networks, making it suitable for resource-constrained environments like mobile devices.

Model	Top-1 Accuracy	Million Mult-Adds	Million Parameters
Inception V3 [18]	84%	5000	23.2
1.0 MobileNet-224	83.3%	569	3.3
0.75 MobileNet-224	81.9%	325	1.9
1.0 MobileNet-192	81.9%	418	3.3
0.75 MobileNet-192	80.5%	239	1.9

Figure 3.6: MobileNet accuracy vs Inception V3 accuracy on Stanford Dogs. Source [22].

3.2.4 EfficientNet

EfficientNet [24] is a convolutional neural networks developed by Google researchers that aims to achieve the state-of-the-art accuracy while being computationally efficient. The key idea is to scale up the model's depth, width, and resolution in a balanced manner, allowing it to perform well across various resource constraints.

Before EfficientNet the common way to scale up neural networks was to scale one of the three dimensions (depth, width and image size) and, despite the fact that scaling up two of the three dimensions arbitrarily was possible, it required tedious manual tuning resulting in sub-optimal accuracy and efficiency.

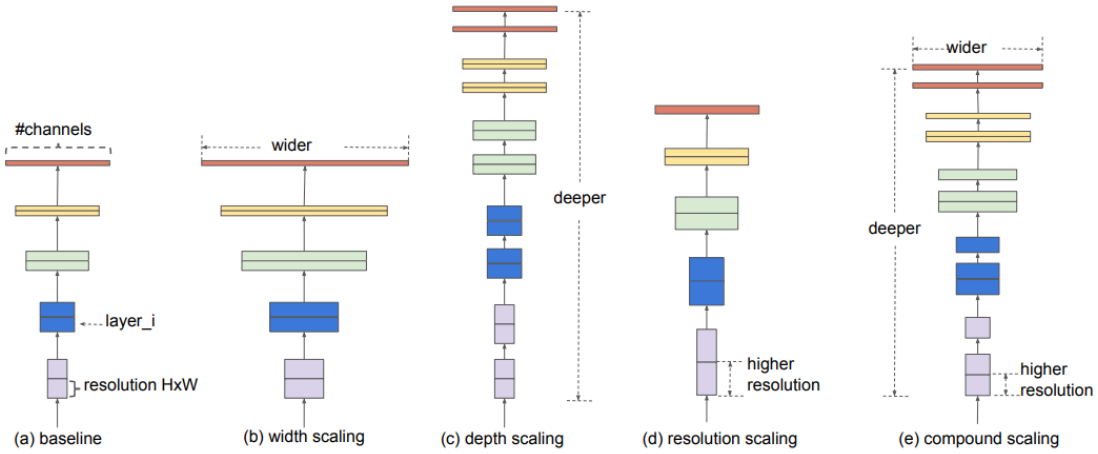


Figure 3.7: Model scaling. Source [24].

The idea of *compound scaling* is the key point of this new approach. EfficientNet uses a compound scaling method that uniformly scales the network depth, width, and resolution with a set of fixed scaling coefficients. This approach ensures that the model's performance improves as resources (such as computational power and memory) increase.

The problem can be formulated as an optimization problem:

$$
\begin{aligned}
&\max_{d,w,r} \text{Accuracy} \\
&\text{s.t. } N(d,w,r) = K \\
&\sum_{i=1}^{s} \hat{F}^d \cdot \hat{L}^i \leq Xh^r \cdot H^i, r \cdot W^i, w \cdot C_i^i \\
&\text{Memory}(N) \leq \text{target memory} \\
&\text{FLOPS}(N) \leq \text{target flops}
\end{aligned}
$$

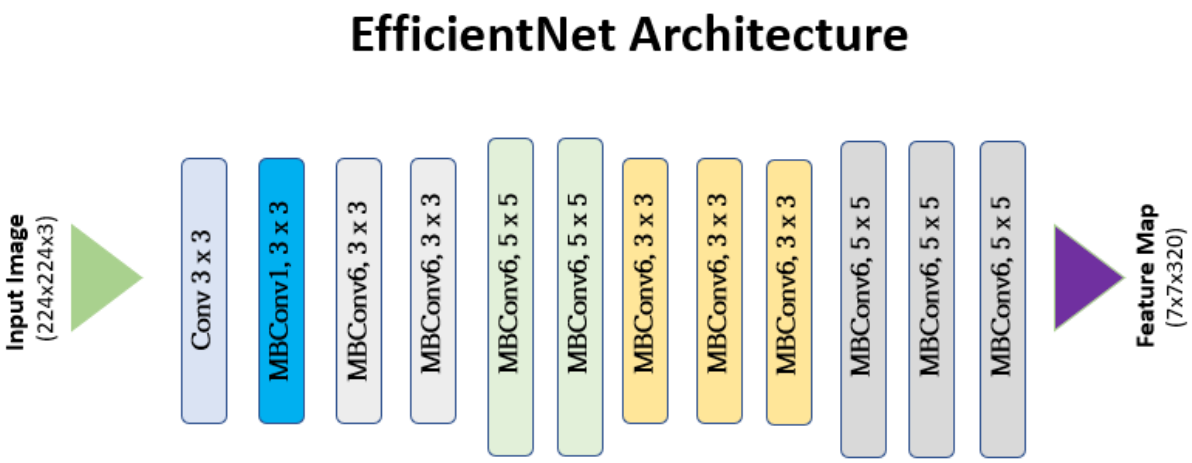


Figure 3.8: EfficientNet architecture

where *w, d, r* are coefficients for scaling network width, depth and resolution, while $\hat{L}^i, \hat{F}^i, \hat{H}^i, \hat{W}^i, \hat{C}^i$ are predefined parameters in baseline network.

The main difficult of the optimization problem is that the optimal *w, d, r* depend on each other and the values change under different resource constraints.

As visible in 3.7 the compound scale simultaneously width, depth and resolution. EfficientNet includes variations of the model optimized for mobile devices, where computational resources are typically more limited. These smaller models are efficient in terms of both model size and computational cost while still maintaining high accuracy.

The baseline network developed and called EfficientNet is built by leveraging a multi-objecive neural architecture search that optimizes both accuracy and FLOPS and is showed in 3.8.

3.3 Previous works

A great deal of work has been devoted, over time, to trying to make face recognition algorithms fairer [27], [28], [29], [30]. Finding the liability of a technology is imperative to contribute in its development in fairness and accountability and useful to understand how deployment can be made in a fair and correct way.

In this context, terrific work like The Perpetual Line-Up [31] are fundamental to highlights the potential biases of face recognition systems that are currently used

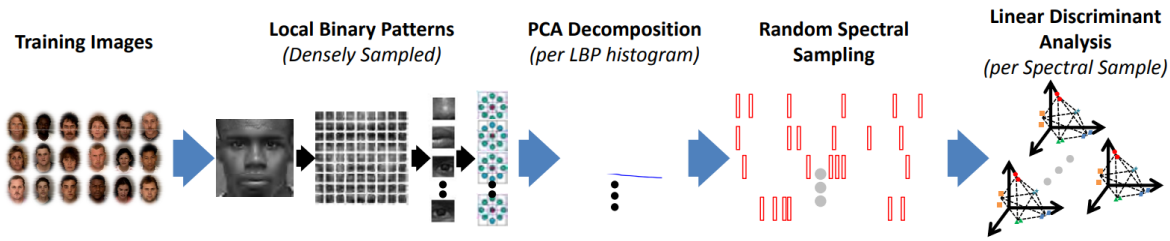


Figure 3.9: Overview of the Spectrally Sampled Structural Subspace Features (4SF) algorithm. This custom algorithm is representative of state of the art methods in face recognition. By changing the demographic distribution of the training sets input into the 4SF algorithm, we are able to analyze the impact the training distribution has on various demographic cohorts. Source [32].

in USA for surveillance systems and crime detection.

Other works focus on biases of race/ethnicity, gender and age. For example Face Recognition Performance: Role of Demographic Information [32] performs an in-depth analysis on the role of demographic information in face recognition performances, underlined the biases of the society reflected by the computer vision models. They underlined that "automated face recognition algorithms are ultimately based on statistical models of the variance between individual faces".

For instance, these algorithms seek to minimize the measured distance between facial images of the same subject, while maximizing the distance between the subject's images and those of the rest of the population. For this reason, if the training set is not representative of the data an algorithm will be operating on, then the performance of the system may deteriorate. The perfect example is the case in which an algorithm is trained on White faces, and after the training is operated on Black faces, the learned representation may discard information useful for discerning Black faces, resulting in a drop of performances. The idea of the researcher of [32] is to train a custom model, namely 4SF, with a structure visibile in figure 3.9, on different subsections of a dataset with a specific focus on a certain ethnicity, on a certain gender or on a certain age group.

The obtained results are visible in figure 3.10 and showed how the algorithms achieved the lowest matching accuracy on Black cohort, even with balanced training. The accuracy is also lower on females than males, and using the principle of intersectionality early explained [2], it is easy to understand how these algorithms struggle to efficiently recognize Black women.

As is easily observable, none of the papers mentioned delves into the question of investigating gender identifications outside the gender binary. Trying to fill the space left empty by the other works a recent work [33] introduced the gender non-conforming individuals in an attempt of bias mitigation. They underlined that

LISTED ARE THE TRUE ACCEPT RATES AT A FIXED FALSE ACCEPT RATE OF 0.1% FOR EACH MATCHER AND DEMOGRAPHIC DATASET.

	Females	**Males**
COTS-A	89.5	94.4
COTS-B	81.6	89.3
COTS-C	70.3	80.9
LBP	54.4	74.0
Gabor	56.0	68.2
4SF trained on All	73.0	86.2
4SF trained on Females	71.5	85.0
4SF trained on Males	69.0	86.3

	Black	**White**	**Hispanic**
COTS-A	88.7	94.4	95.7
COTS-B	81.3	89.0	90.7
COTS-C	74.0	79.8	87.3
LBP	65.3	70.5	73.5
Gabor	61.6	63.7	70.9
4SF trained on All	78.4	83.0	86.3
4SF trained on Black	80.2	81.0	59.8
4SF trained on White	75.4	84.5	59.9
4SF trained on Hispanic	74.5	80.2	60.1

	18 to 30 y.o.	**30 to 50 y.o.**	**50 to 70 y.o.**
COTS-A	91.7	94.6	94.4
COTS-B	86.1	89.1	87.5
COTS-C	76.5	80.7	83.6
LBP	69.4	74.7	75.1
Gabor	61.7	68.2	65.7
4SF trained on All	81.5	85.6	83.6
4SF trained on 18 to 30 y.o.	83.3	85.9	80.7
4SF trained on 30 to 50 y.o.	82.1	86.0	82.2
4SF trained on 50 to 70 y.o.	78.7	84.5	82.0

Figure 3.10: Results of the analysis conducted using 4SF. Source [32].

the lack of representation of the LGBTQIA+ population in current benchmark databases leads to a false sense of universal progress on gender classification and facial recognition tasks in machine learning. Their baseline model failed to predict gender non-conforming individuals, so their work was focused on the creation of an inclusive dataset to train their model. In this way they reached decent accuracy on the category, distributed as showed in Figure 3.11.

However, for the previously expressed concept of intersectionality, the performances of the algorithm on subcategories of the datasets (for instance, Black gender non-conforming individuals) still struggle if compared to the performances obtained on White males.

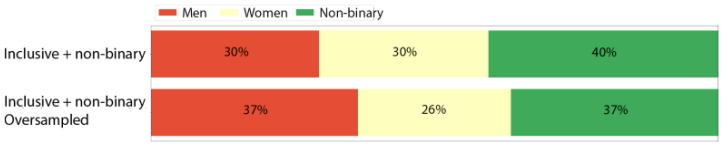


Figure 3.11: Distribution of images in the dataset. Source [33].

Chapter 4

Methodology

This chapter is dedicated to an in-depth look at our technology, trying to explain in as much depth as possible the rationale behind the choices made and the methodology used to reach the objective of the research. Summarizing, the ultimate goal of this chapter is to exploit the basis introduced in the previous chapter to have a deep understand of the decision-making process through a comprehensive analysis.

The first section is dedicated to the challenges proposed in the creation of the dataset and the realization of our final dataset. The second section is dedicated to the comparative analysis of existing technologies. The third and last one is dedicated to LIME (Local Interpretable Model-Agnostic Explanation) [34] and to the approach used to analyze the obtained results.

4.1 Dataset

In order to understand the criticalities of the proposed work it is necessary to introduce the dataset used in the final work, keeping in consideration the difficulties of collecting pertinent and well-labeled data.

The starting point of the work is CelebA [35]. CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes with more than 200k celebrity images, with an enormous number of attribute annotations. The images cover large pose variations and differences in backgrounds, creating large diversities and quantities, including 10,177 different identities and a total number of 202,599 face images as visible in Figure 4.1.

CelebA, with the amount of images and annotations associated, is a perfect starting point for training every model which purpose is a face recognition task or sub-task of face recognition.

For the purpose of this research, the biggest problem of CelebA is that the

Figure 4.1: Sample images from CelebA. Source [35].

annotation about gender is codified as "Male" with a binary annotation (0 for females, 1 for males). This is a huge obstacle in the proposed research and is a clear symptom of the problem of invisibility explained in-depth in 2.

Wikipedia offers a list of people with non-binary gender identities [36]. The list is a collection of notable people who identify with a gender outside the gender binary. Thanks to the work of Wikipedia a previous research [33] created a dataset with 2k images of 67 different gender non-conforming individuals, make it publicly available.

To ensure a balanced representation of gender, the dataset of gender non-conforming individuals was augmented by adding 4,000 images, evenly distributed between males and females, all of different individuals, randomly selected from CelebA. This augmentation resulted in a collection of 6,000 diverse images, balanced according to gender, as shown in figure 3.11.

Despite the possibility to create an imbalanced dataset having available 200,000 images from CelebA, a balanced data was preferred. The reasons behind this choice are numerous and various and the fragility of CNN on unbalanced dataset is discussed in this paper [37].

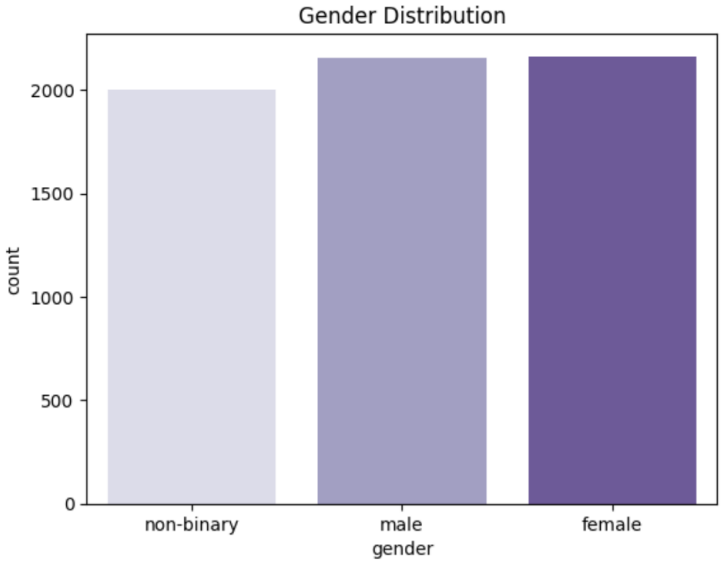


Figure 4.2: Distribution of the initial dataset

The first reason is to avoid bias, imbalanced datasets can introduce bias into the model and in a such complex task the classification on the gender non-conforming class would probably have been very low if this class had been sacrificed in training. Looking at the work [33] clear suggests that it would have been even more likely to improve the total classification by focusing model training on the gender non-conforming class, but a balanced dataset is a good compromise in this case. The second reason is the attempt to reach an improved generalization on unseen data, while the objective of a balanced dataset is to provide to the model sufficient examples from each class. The third reason is the attempt to avoid overfitting, because with an unbalanced representation of classes the model could learn to memorize the majority class rather than capturing meaningful patterns. The last, but not least reason is the attempt to reach fairness and inclusivity. A balanced dataset ensures that the model is trained on an equal representation of different groups, promoting fairness in the model's predictions. This could help, in the case presented here, to prevent the marginalization of gender non-conforming individuals.

The dataset introduced, however, presents several problems related with the small amount of data available. The problem of data availability and, as a consequence, data diversity is crucial and could have a huge impact on the final results and on the possibility to use the trained model in a real-world scenario and in future and more complex application. The need to expand the dataset, however, clashes with the lack of available data, the result of the aforementioned problem of invisibility. In this difficult terrain the idea is to contribute to expand the dataset of gender

non-conforming individuals created by [33]. Merging the aforementioned dataset with last version of the list of non-binary celebrities offered by Wikipedia [36] it was possible to extract a list of names whose images were not yet present in the dataset. Starting from this list the idea was to search all the names on social network, in particular Instagram, to extract and collect as many images as possible for each of them.

Thanks to this time-consuming work the availability of images for the gender non-conforming class went from 67 to 118 different individuals and from 2000 single images to 3843 single images. The new distribution of gender in the dataset and a comparison with the initial dataset is visible in 4.3.

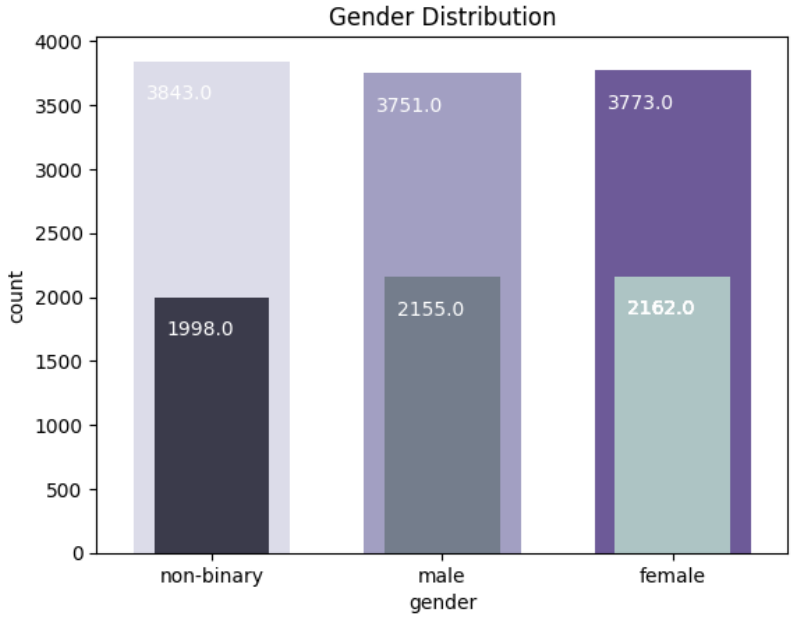


Figure 4.3: Gender distribution of the final dataset compared with gender distribution of the initial dataset

Although the number of different individuals is still not satisfactory it could be considered a good starting point for future research in this field. An example of images present in the final dataset is visible in figure 4.4 and are portraits and close-ups of celebrities in various contexts, starting from self pictures taken with the smartphones, to pictures taken on the red carpet to picture taken by fashion shooting or magazines.

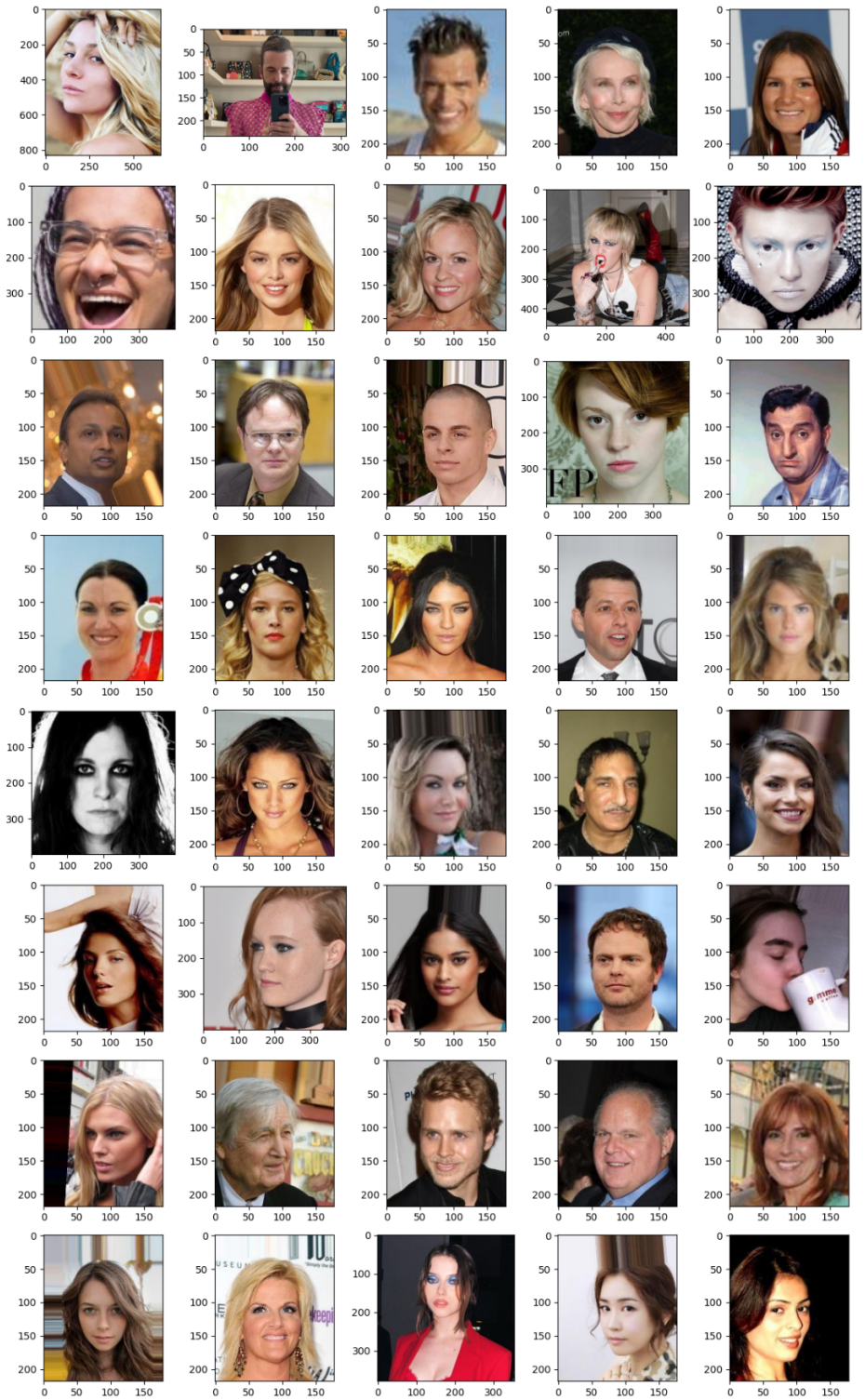


Figure 4.4: Example of images of the final dataset

4.2 Proposed Pipeline

This section serves as a detailed exposition of the adopted pipeline which is essential to better understand and replicate the research in the future. The complete pipeline, visible in 4.5, describes a systematic approach comprising various stages, each crafted to address the challenges of gender recognition trying to prioritize fairness.

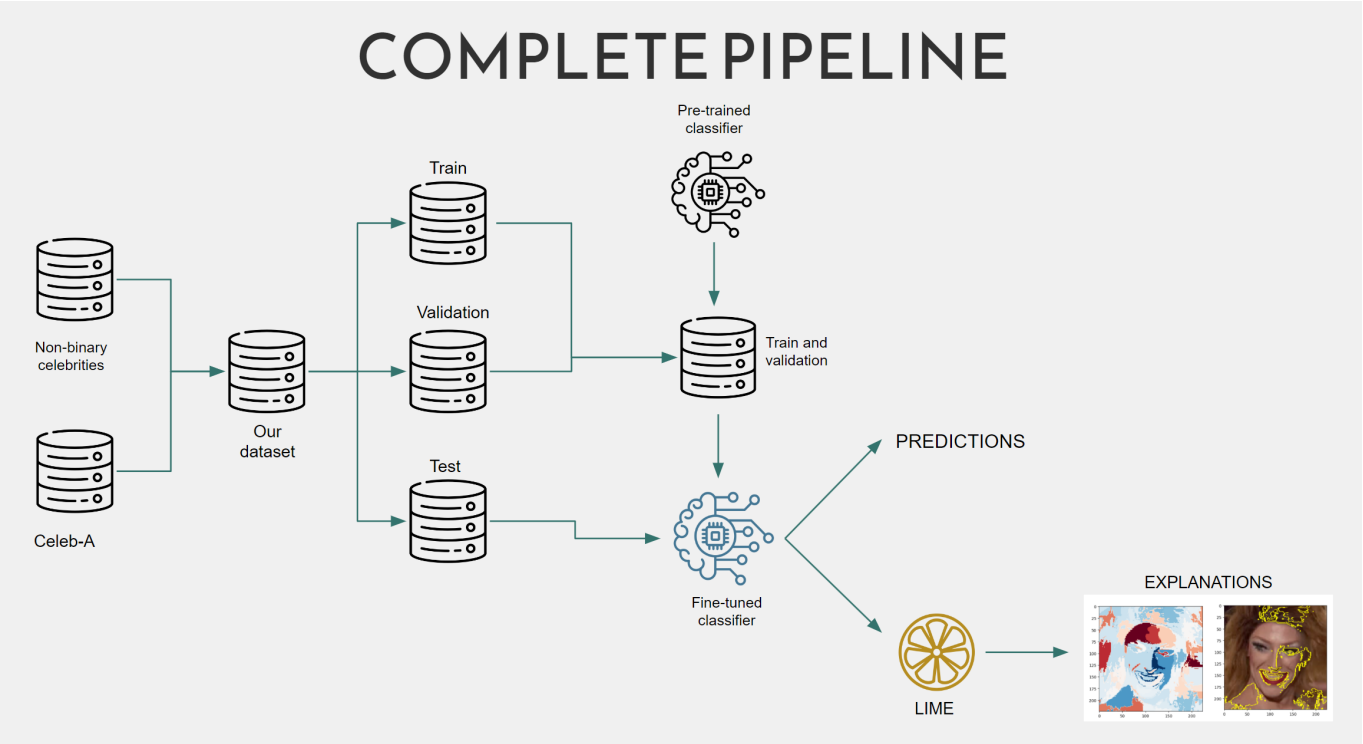


Figure 4.5: Complete pipeline adopted

The implementation of the pipeline begins with the data acquisition through different stages, in order to obtain the dataset described in the dedicated section. The dataset was then divided into train, test and validation in order to perform the fine-tuning of different computer vision models to obtain the fine-tuned classifier. The resulting classifier is then used to produce quantitative results. Simultaneously, LIME [34] is used so as to obtain qualitative results, which analysis is very useful for the ultimate purpose of the research.

After this brief recap of the complete pipeline let us go into more detail by better analyzing the various sections of it.

4.2.1 Train, Test and Validation

Splitting a dataset into train, test, and validation sets is a fundamental practice in machine learning to develop and evaluate models effectively. Initially, the dataset is divided into three distinct subsets: the training set, the validation set, and the test set. The training set, comprising the majority of the dataset is used to train the machine learning model by exposing it to labeled examples.

The validation set serves as a means to fine-tune the model's hyperparameters and monitor its performance during training. It helps in preventing overfitting by providing an independent dataset for evaluation. The test set remains untouched during the training phase and is used to assess the model's performance after training. It ensures an unbiased evaluation of the model's generalization ability to unseen data and provides insights into its real-world performance.

Due to the special nature of the dataset that was constructed for this research, a particular choice was made regarding the division into train, test, and validation. The number of different individuals labeled as gender non-conforming is 118 so that there is no repetition of the same individual in both the train set and the test set, for the gender non-conforming subset of the dataset the division was made on the single individuals and not on the totality of the images.

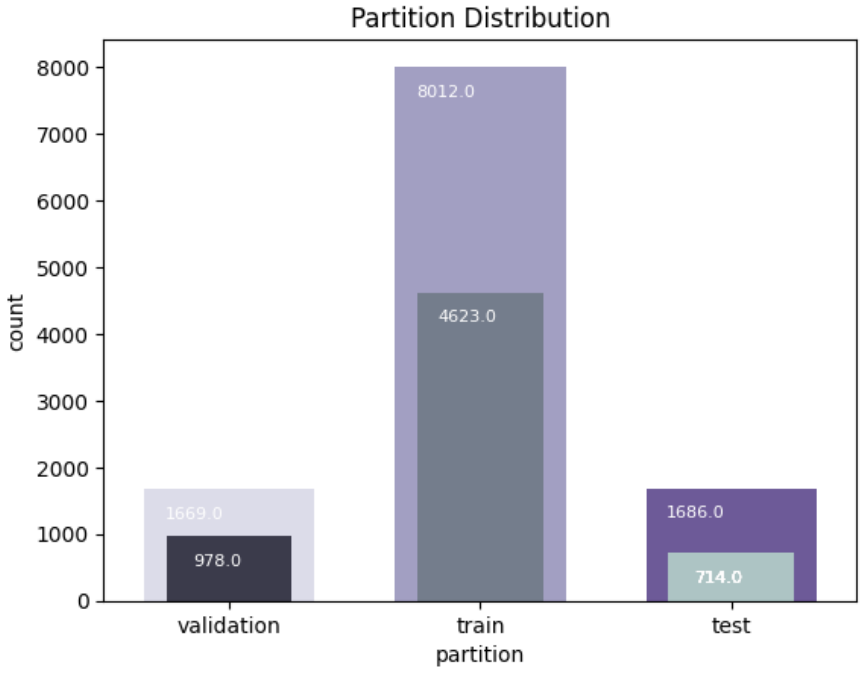


Figure 4.6: Train, test and validation splits of the final datasets

For instance, having 20 different individuals, namely *individual_1* to *individual_20*, with a number of images for each ranging from 10 to 70, with a 70/15/15

split individuals from *individual_1* to *individual_14* are selected for the training set, *individual_15*, *individual_16* and *individual_17* for the validation set and the remaining for the test set. This method does not ensure that we have 70% of the images of gender non-conforming individuals in the train split, but it does ensure that there are no repeats of the same individual in the train split and the test split. Instead, for the images labeled as male/female the 70/15/15 split is performed on the totality of the images because each image represents a different individual. The resulting split is visible in 4.6.

4.2.2 Fine-tuning of pre-trained models

The pre-trained model chosen for this work are ResNet [21], MobileNet [22], EfficientNet [24] and Inception [23] and a deep analysis of each of them is available in section 3.2.

Due to the limited amount of data available, particularly on the gender non-conforming class, was taken the decision to perform a k-fold cross validation during the training phase through the different cross validation criterion [38].

K-fold cross-validation is a robust technique used to evaluate machine learning models by dividing the dataset into k equal-sized subsets or folds. In each iteration of the cross-validation process, one fold is used as the validation set, while the remaining k-1 folds are used for training the model. This process is repeated k times, with each fold serving as the validation set exactly once. By rotating the validation set across different subsets, k-fold cross-validation ensures that the model is trained and evaluated on different combinations of data, thus providing a more reliable estimate of its performance.

After completing all k iterations, the performance metrics obtained from each fold, such as accuracy or F1 score, are averaged to derive a single evaluation metric for the model. This aggregated metric reflects the model's overall performance across the entire dataset and helps in assessing its generalization ability to unseen data. K-fold cross-validation is particularly beneficial for detecting overfitting and variance in the model's performance, as it reduces the dependency on a single train-test split and maximizes the utilization of available data for training and validation.

Summarizing, k-fold cross-validation is a versatile technique that enhances the reliability of model evaluation and provides valuable insights into the model's behavior across different subsets of the data, contributing to the creation of a more robust model with a limited set of data.

Given the size of our database, we opted to use the transfer learning technique. It consists in applying a model pre-trained on a large-scale database (in our case, ImageNet [39]) or pretrained for a different task to a different but related task. The objectives of this operation are better results and faster training.

To optimize the training phase's efficacy, several preprocessing operations were applied to the input data. Initially, the pixel values underwent rescaling, a fundamental step aimed at ensuring numerical stability, by constraining them to a normalized range of [0, 1]. This normalization process serves to standardize the data and facilitate convergence during model training.

Subsequently, a series of geometric transformations were implemented to augment the dataset's diversity and enhance its robustness. The first of these transformations involved applying a shear transformation with a configurable range of 0.2. This operation introduces controlled distortions to the images, thereby increasing the dataset's variability and enabling the model to learn to generalize better across different perspectives and orientations.

Following the shear transformation, a zoom operation was applied with a range of 0.2. This zooming technique allows the model to encounter variations in scale, further enriching its understanding of the input images and improving its ability to recognize and classify objects across different magnifications.

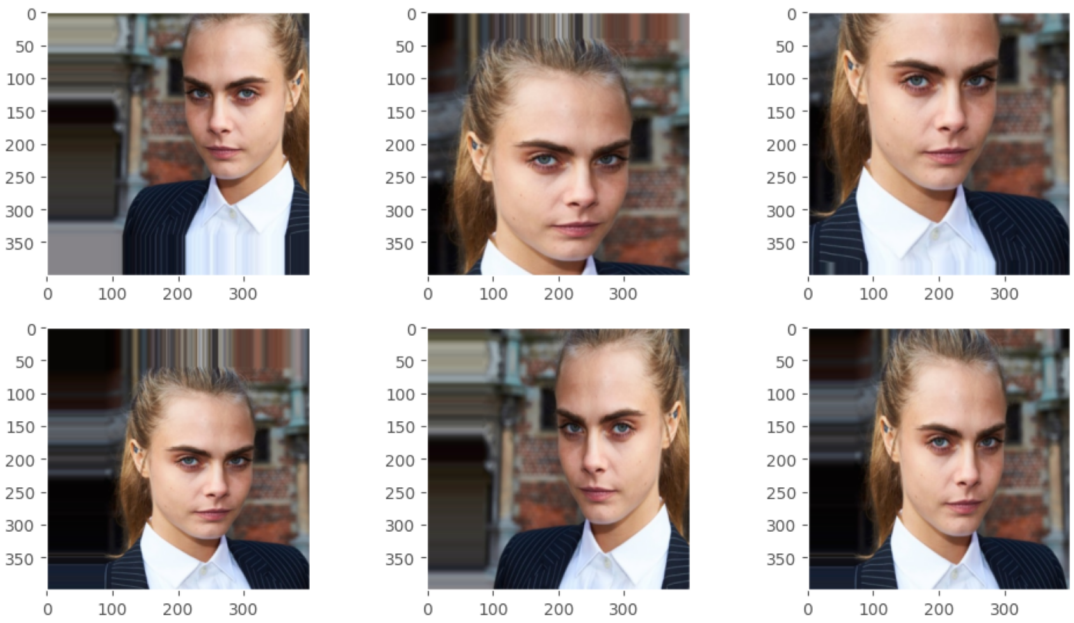


Figure 4.7: Train, test and validation splits of the final datasets

Lastly, horizontal flipping was employed as a final augmentation step. This mirroring technique horizontally flips the images, effectively doubling the dataset size while also providing the model with exposure to mirrored representations of the objects, which aids in promoting robust feature extraction and classification.

The comprehensive data augmentation process is visually depicted in Figure 4.7,

illustrating the progressive transformations applied to the input images. Collectively, these preprocessing techniques play a pivotal role in diversifying the training dataset, empowering the model to generalize better and achieve superior performance across various real-world scenarios.

Utilizing transfer learning, the deep learning model was initiated by leveraging various architectures previously trained on the ImageNet dataset. To tailor the pre-trained model to the specific task, a strategic approach was employed wherein the last 5 layers were frozen, rendering them untrainable. This action enabled the model to preserve the overarching features gained from the extensive ImageNet training set.

Following the freezing of the final layers, custom layers were integrated into the model architecture. Initially, a flattening layer was introduced to convert the multi-dimensional feature maps into a one-dimensional array, facilitating integration with subsequent layers. This transformation lay down the basis for the efficient extraction of features from the input data.

Subsequently, two fully connected layers were added to the model. The first fully connected layer comprehends 1024 neurons, employing Rectified Linear Unit (ReLU) activation function. This layer serves as a feature extractor, enabling the model to discern intricate patterns and relationships within the data.

The final layer of the model comprises 3 neurons, incorporating softmax activation. Tailored to the specific task, this layer facilitates the classification of images into three distinct categories, aligning with the objectives of the project.

Parameter	**Value**
Epochs	5
Batch Size	64
Optimizer	adam
Loss	categorical crossentropy
Image Size	(224, 224)

Table 4.1: Hyperparameters for transfer learning

The training of the various models was performed using the hyperparameters shown in table 4.1. Adam was adopted as optimizer, because it is known for its effectiveness in optimizing deep neural networks and helps to achieve faster convergence, while categorical crossentropy is a commonly loss function employed for multiclass classification tasks. The idea behind the choice of the small number of epochs and the image size, however, lies in the trade-off between computational speed and accuracy.

4.2.3 Explainable AI

The ultimate step of the proposed pipeline is the analysis of the obtained predictions through algorithms of explainable AI.

Typically AI is a blackbox, as showed in 4.8, which given inputs we can only know the outputs, but not the precise logic that led to the results. In many applications, in order to ensure transparency and fairness, is crucial to obtain an explanation of how a certain result is obtained. In this context explainable AI algorithms born.

Explainable AI (XAI) refers to the capability of artificial intelligence systems to provide understandable explanations for their decisions and actions to users. XAI aims to untangle the inner works of artificial intelligence models so as to allow humans to comprehend the motivations behind algorithmic outcomes.

In addition to create a sort of trust on singular predictions, it's imperative to evaluate the model as a whole. Before deploying the model "in the wild" it is necessary to be confident that the model will perform well in the real-world.

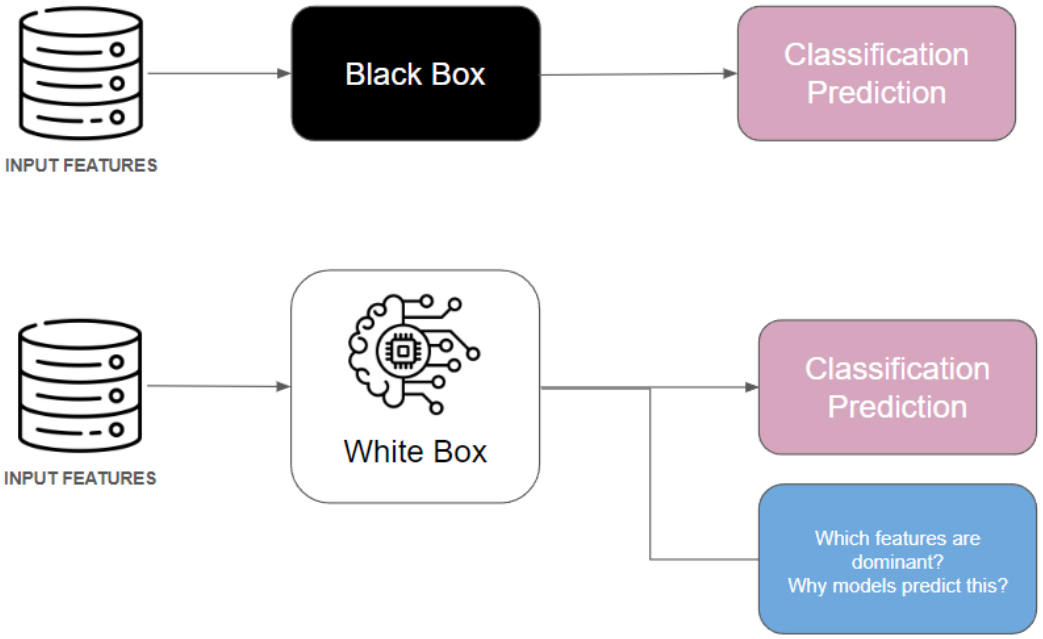


Figure 4.8: Artificial Intelligence vs Explainable AI

Through the various explainable AI algorithms the one chosen to provide the explanation in this research is LIME (Local Interpretable Model-Agnostic Explanations) [34]. LIME is a technique that approximates any black box machine

learning model with a local, interpretable model to explain each individual prediction. The authors of LIME propose the objective of "identify an interpretable model over the interpretable representation that is locally faithful to the classifier" [34]. *Interpretable* represents the will of the authors to use a representation that is understandable to humans whatever the representation of the features within the model is.

Let the model be denoted $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Let the proximity measure between an instance z to an instance x be denoted by $\pi_x(z)$, so as to define locality around x. Let $L(f, g, \pi_x)$ be a measure of how unfaithful g is in approximating f in the locality defined by π_x. Let $\Omega(g)$ be a measure of complexity of the explanation g.

To ensure the interpretability and local fidelity the objective of LIME is to minimize $L(f, g, \pi_x)$, while having $\Omega(g)$ be low enough to be interpretable by humans. So the final explanation of LIME is produced by:

$$\xi(x) = \arg\min_{g \in G} \left(L(f, g, \pi_x) + \Omega(g) \right)$$

To clearly summarize, the idea behind LIME is to perturb the original data points, feed the black box model and observe the corresponding outpus. Then weights those new data points as a *function of proximity* to the original point. Finally the idea is to fit a proxy model (i.e. a linear regression) on the dataset with variations using those sample weights. With this method, summarized in figure 4.9, it is possible to explain each original data with the new explanation model.

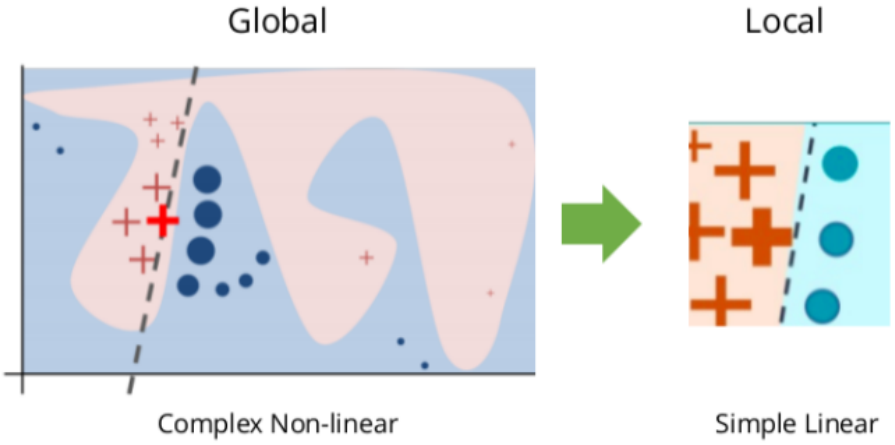


Figure 4.9: Local interpretation produced by LIME vs global non-linear representation of the classification. Source [34].

Thanks to the use of LIME it was possible to create heatmaps useful to understand and analyze which features were the most significant for the classification of each model and for each class, providing a useful tool to evaluate if the premise of *gender as an act of performance*, explained well in 2.4.2, is confirmed by the classification of computer vision models.

Chapter 5

Results and discussions

This chapter delves into the key findings of our research on Automatic Gender Recognition on three classes which methodology is explained in 4, exploring both quantitative and qualitative aspects through the lens of mathematical metrics (i.e. accuracy, f1 score) and LIME explanations. By combining these perspectives, we aim to gain a comprehensive understanding of how the proposed models perform, how the models make their predictions and the social implications of this technology.

The quantitative analysis revealed that it is possible to obtain sufficient results on this task. Additionally, LIME provided valuable qualitative insights into the patterns that define, through multiple lenses of social culture, the performed gender of an individual.

The objective of this chapter is to better analyze the results and perform a critical analysis of them to gain awareness of the proposed topic and understand if it can be effectively translated into a real world scenario. The focus of the last section of this chapter is wondering what are the implications of technology and for what purposes can it be used ethically without negatively impacting the life of the individual.

Summarizing, this chapter is divided in three section: *quantitative results*, *qualitative results* and *discussions*.

5.1 Quantitative Results

Firstly, it is necessary to introduce the metrics used to perform the quantitative evaluations of the models. The accuracy is the simplest metrics used to evaluate the performance of a machine learning model. It measures the proportion of correct predictions made by the model out of the total number of predictions made. The

formula of accuracy is:

$$Accuracy = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Samples}}$$

To evaluate the classification of the various classes is interesting to observe the F1-score, defined as:

$$F_1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision represents the ratio of true positives to all positive predictions, while recall (or sensitivity) represents the ratio of true positives to all actual positives in the dataset.

	ACCURACY	**F1 SCORE**		
MODELS	**Overall**	**Male**	**Female**	**Non Binary**
ResNet50	68.8%	0.47	0.70	0.84
MobileNetV2	64.6%	0.48	0.71	0.68
InceptionV3	92.2%	0.96	0.91	0.90
InceptionV3	83.2%	0.82	0.80	0.82
EfficientNetB0	94.4%	0.92	0.93	0.99
EfficientNetB4	94.3%	0.91	0.94	0.98

Table 5.1: Accuracy and F1-score of the different models. In red are showed the results obtained on the final dataset, in black the results obtained in the smallest dataset.

The results are visible in 5.1 and it is necessary to observe some important evidences. The first evidence is that different models perform in a really different ways on this particular task using the methodology explained in 4. ResNet50 [21] and MobileNetV2 [22] obtained the worst results. It is interesting to observe that, in the results of ResNet50, the F1-score of non-binary class is far higher than the F1-score obtained on the other classes. This outlier is most likely due to the small amount of different individuals in the non-binary class that could results in a slight overfitting over this class. It is clear that *InceptionV3* and *EfficientNet* outperformed the other models. Especially EfficientNet seems to perform well on both datasets with this training strategy.

It is necessary to open a brief aside on the training methods to better assess why certain models achieved certain results. Due to the limitations of the training environment it was necessary to choose models with a good tradeoff between computational complexity and performances. EfficientNet seems to be the best

suited one keeping in consideration the performances and the ability to be trained on a small amount of data with a decent computational complexity. Observing the results of InceptionV3 it is visible how the model trained on the first dataset, smaller than the final dataset, obtained better results than the model trained on the final dataset. This is probably due to the fact the adopted strategy designed to ensure a reasonable training time.

Figure 5.1: Confusion matrix of the results obtained with EfficientNetB0 on the first dataset and with EfficientNetB4 on the final dataset

Taking advantage of better hardware, it would be interesting to increase the number of epochs within each iteration of the k-fold cross validation, better tuning the hyperparameters to ensure a well-suited training for each model. Under the current conditions, using the same strategy for all models, looking at the results there seem to be some models that are in danger of overfitting and others that do not have enough data to properly "learn" to recognize and classify the various images.

5.2 Qualitative Results

To comprehend the fundamental distinctions among the models is necessary to evaluate the results obtained through the use of explainable AI algorithms (i.e LIME [34]), described in 4.2.3. This section is fundamental to gain insights into the rationale behind the classification of images, enabling us to discern why a particular classification is assigned and identify the facial features and components that carry greater significance in the classification process.

The heatmaps depicted in figure 5.2 serve as an entry point for comprehending how different features contribute to the classification of images by various models. In particular examining ResNet (the first image on the left in the figure) it is

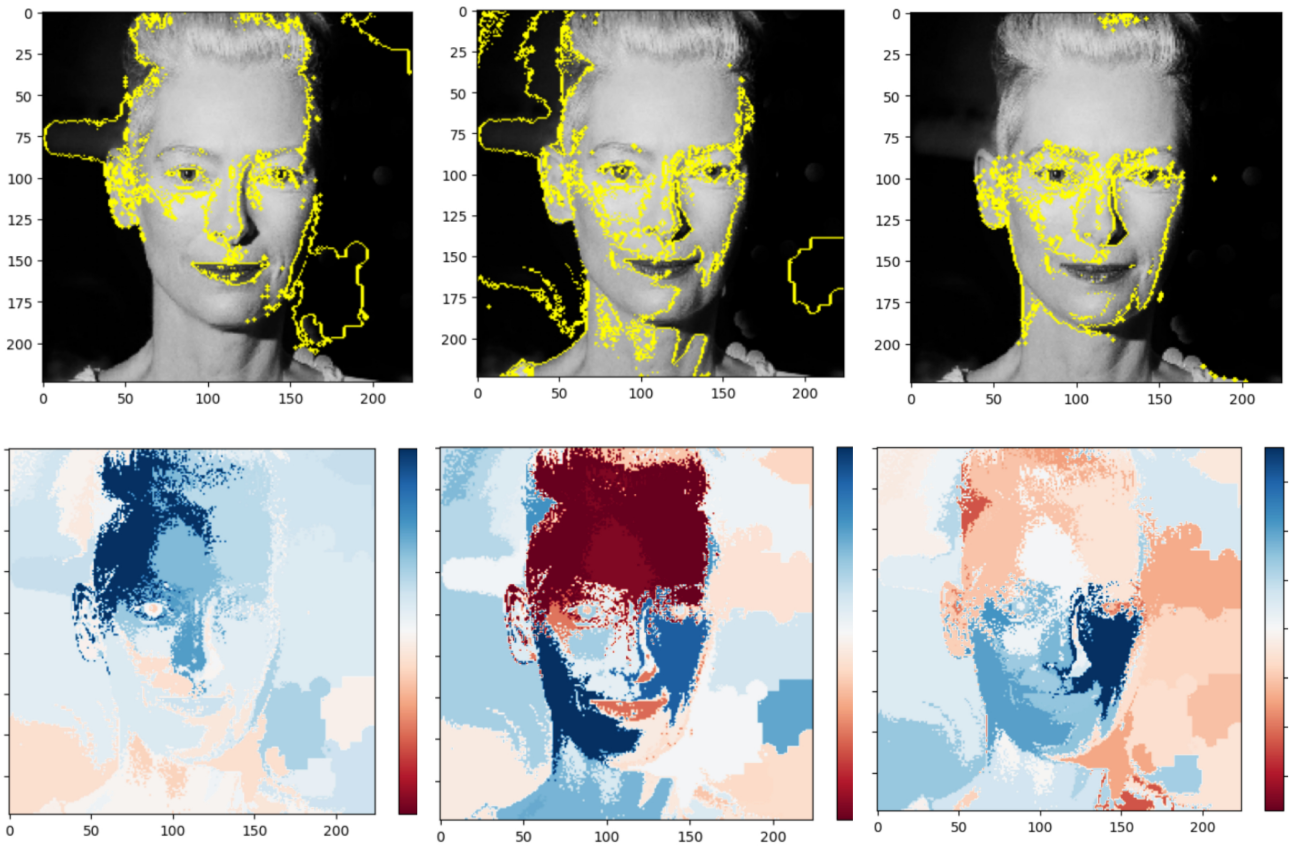


Figure 5.2: From left to right: ResNet50, InceptionV3 and EfficientNet classification. The bluest areas in the heatmaps are the most important for the classification.

evident that it places significant importance, in the classification of this particular image, on the region of the forehead and the initial portion of the hair. From the quantitative results in table 5.1 it is clear that the approach of ResNet is probably not the best one. Conversely, when the attention is turned to InceptionV3 and EfficienNet models, which happen to be the most accurate models, emerges a distinct pattern. These models exhibit little to no emphasis on the forehead region, instead their focus is predominantly directed towards cheeks, cheekbones, eyes, nose and lips. These facial features prioritized by EfficientNet and InceptionV3 intuitively seems to be more descriptive and cross-referencing the quantitative and qualitative results seem to confirm the sensation.

In essence, this comparison underscores how different deep learning models prioritize and leverage distinct facial regions for accurate classification. Observing

that EfficientNet is the most performative model in terms of quantitative results from this point the qualitative results that will be analyzed are the one produced by EfficientNet.

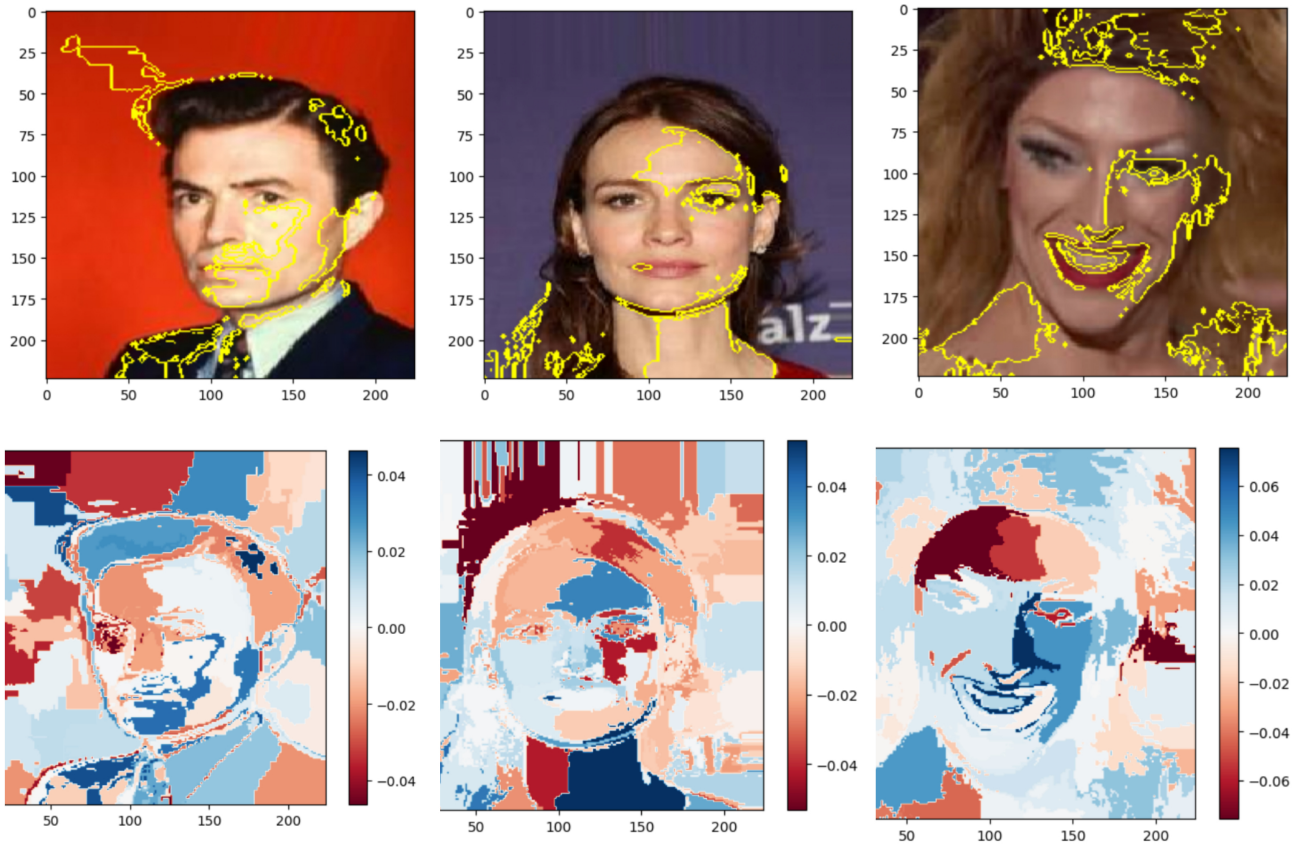


Figure 5.3: From left to right: which features are used by EfficientNet to classify males, females and gender non-conforming individuals. The bluest areas in the heatmaps are the most important in the classification.

The figure 5.3 depicts the differences regarding the region of the face used by EfficientNet in the classification of males, females and gender non-conforming individuals. The first impression is that for classify a male the model focuses on the jawline and cheekbones. These features appear to be key determinants in the gender recognition process for this class. For the female class the model seems to be influenced by the forehead, by the neck and by the hair. Probably it focuses on the skin of the mentioned regions. For the gender non-conforming class the model appears to be guided by the region of the lips and eyes, especially if is present a certain kind of make-up.

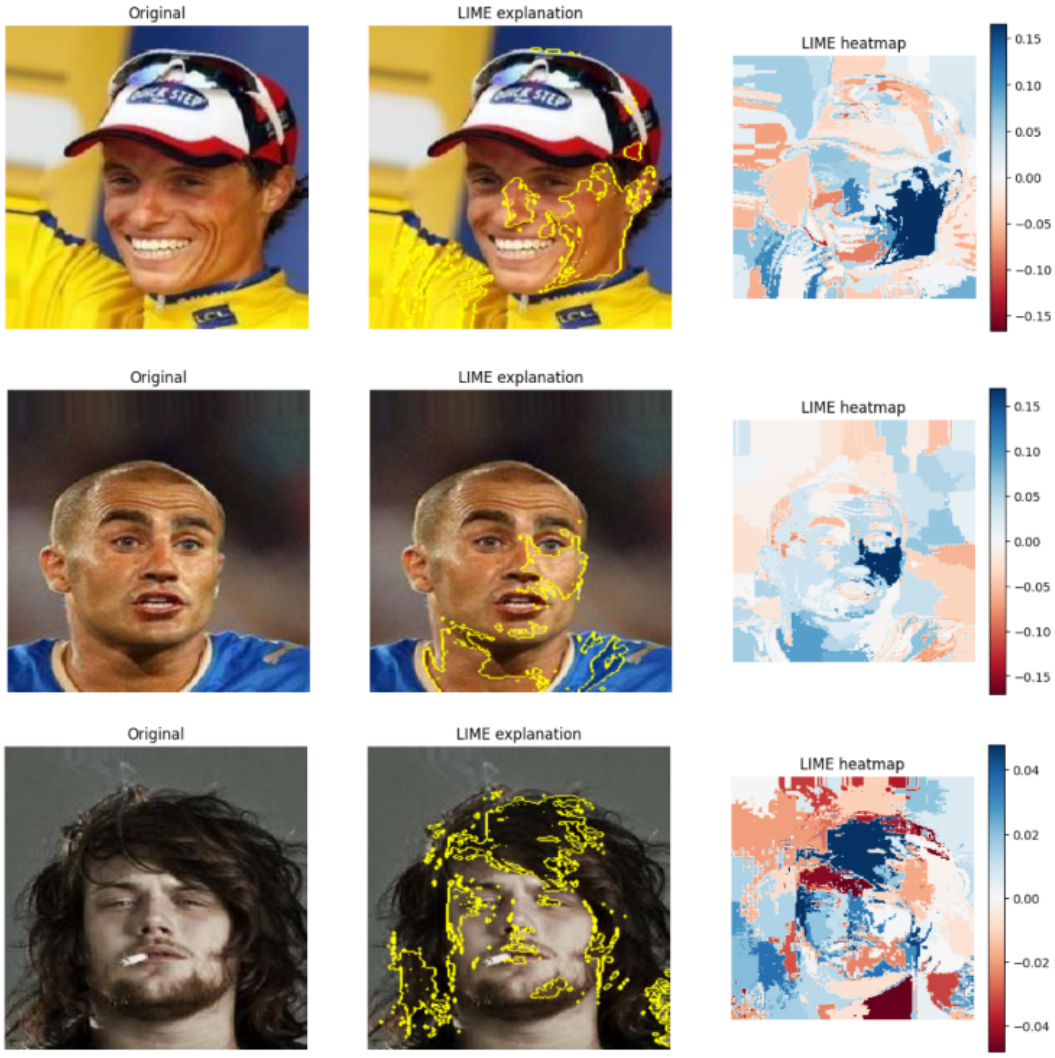


Figure 5.4: 3 examples of well classified males with the explanations produced by LIME.

Focusing on the male class other examples visible in Figure 5.4 seem to confirm the first impression taken from the Figure 5.3. For the first man the area of the jawline is crucial for the classification, for the second one the region of the cheekbones appear to be the more discriminative and for the last one the focus is shifted on the region of the beard and mustaches.

Focusing on the female class the examples visible in Figure 5.5 show that, with a good probability, the skin of certain region of the faces is essential for the classification. As visible in the last two images the region of the forehead

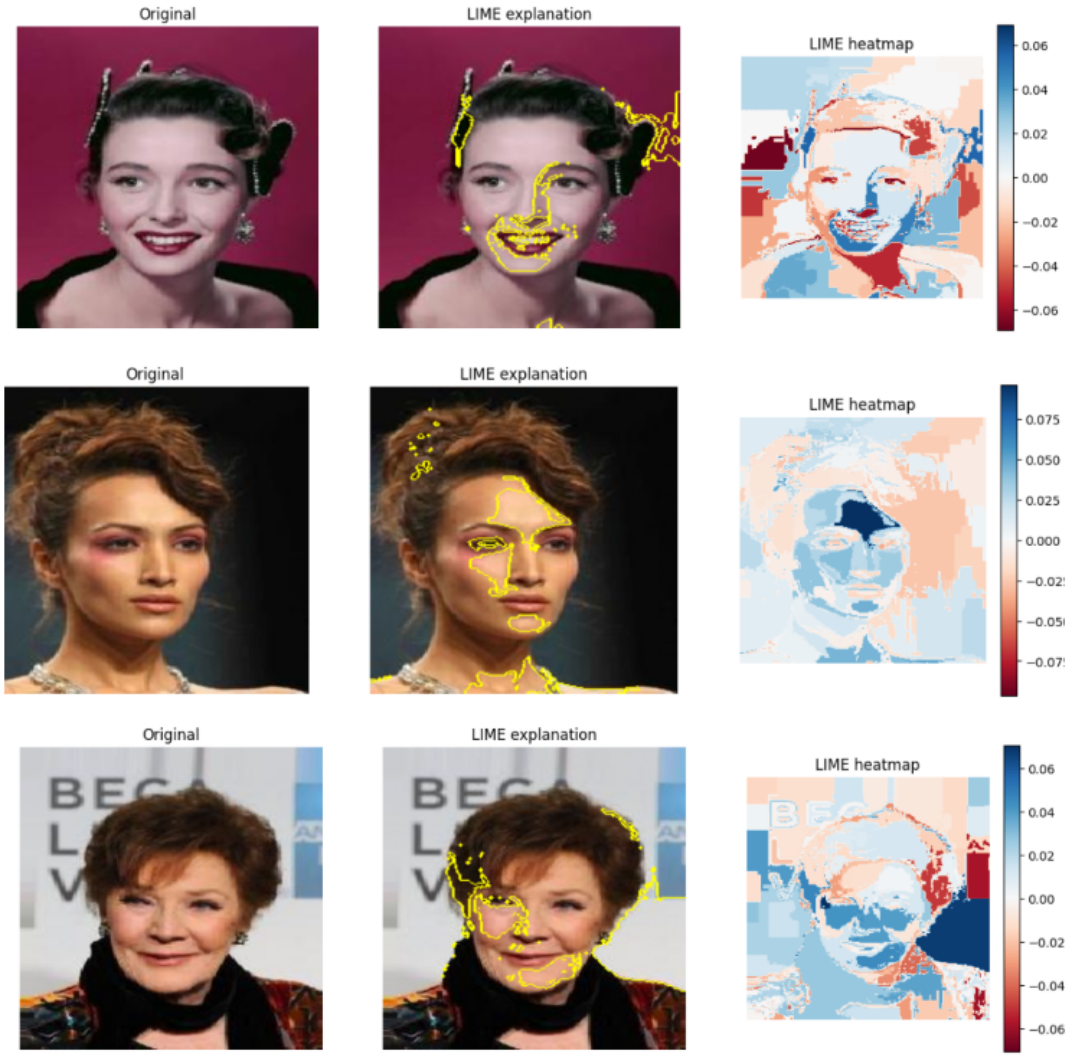


Figure 5.5: 3 examples of well classified females with the explanations produced by LIME.

and the cheeks under the eye seem to be crucial. Also the area of the chin is kept in consideration to distinguish this class, instead, lips are not given special consideration. Counter-intuitively, hair also seems not to be particularly important for this model because, even when it is taken into account, it is not among the key features for classification.

Starting from the gender non-conforming individual present in figure 5.3 we can observe that make-up lips and make-up eyes seem to be important features for the classification of this class. The last row of the figure 5.6 seems to confirm the

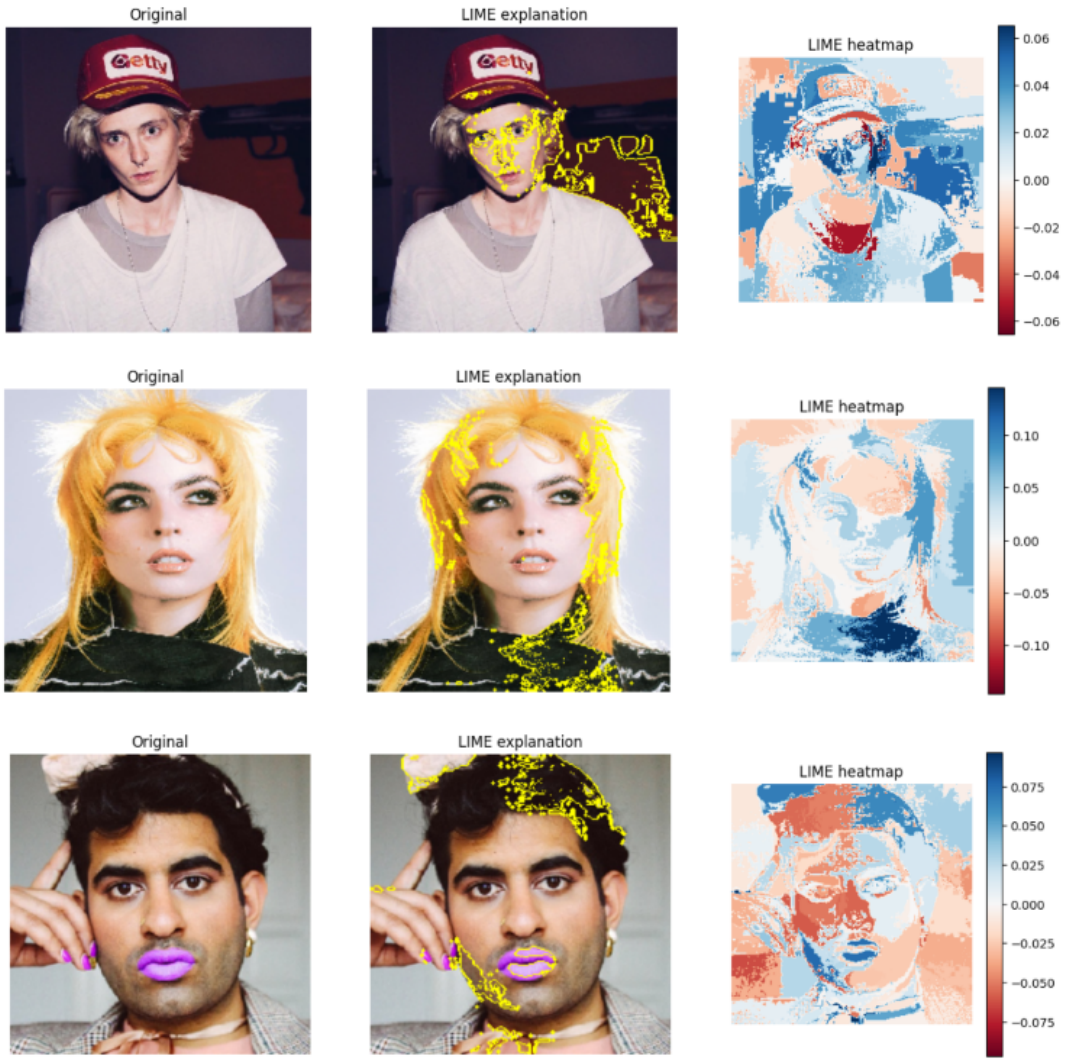


Figure 5.6: 3 examples of well classified gender non-conforming individuals with the explanations produced by LIME.

first impression, because the make-up lips are the most important region for the classification of this image. An important focus is also present on the hair and it is possible to imagine that particular colors or haircuts present in the training dataset may have contributed to the model's adoption of this feature.

5.3 Discussions

Taking into consideration what was discussed in the chapter 2 and the analysis of the qualitative results the amount of material to discuss is certainly significant.

A brief recap: Judith Butler redefines traditional notions of gender by proposing that is not an inherent or fixed attribute but a social construct that individuals continually enact and express through their actions, language, and behavior. This theory suggests that people perform their gender roles and that these performances are integral to the construction of one's gender identity. Starting from the definition of gender as an historical situation, theorized by Simone de Beauvoir [16], she emphasizes the role of societal norms in reinforcing or challenging gender expectations and allows a more fluid understanding of gender, transcending binary categories. This concept has had a profound influence on queer theory and discussions of non-binary and transgender identities, offering a more inclusive and dynamic perspective on the nature of gender. The same concept is carried forward by West and Zimmerman in their work *"Doing Gender"* [17]. In a similar way they challenge the notion of gender as a fixed and inherent characteristic, emphasizing that it is a social process enacted and reinforced through everyday interactions. Understanding gender in this way partly explains why certain types of facial features, not necessarily related to biological sex, are very important in trying to understand a person's gender.

Keeping in considerations the theory just explained, the concept of gender as an historical situation and gender as performative act it is possible to look at the results obtained on the various classes.

The analysis over the gender non-conforming class is interesting to confirm or disprove the premise of this work: could the gender be considered as an act of performance to perform automatic gender recognition tasks? In a nutshell, it is possible to answer that interpretation of these heatmaps may rely on the concept of gender as *performative act* [18], because some regions of the faces can be modified, with make-up for example, to represent and communicate something to those who look at us. Lips, hair, eyes, but also piercings or tattoos could communicate what an individual want to communicate to the world. Observing the figure 5.7 results clear how clothing can influence the algorithm. Glasses of a certain shape carry with them a meaning, the same for accessories, hats, style of clothes and everything that an individual can use to communicate something in a certain culture.

The culture is fundamental in this analysis. If the premise is that the expression of gender is built during the history of an individual in a certain environment results clear that the environment plays a key role in the final expression of an individual. For this reason, to think of extrapolating the model presented in this research, using the same dataset, for use in other cultures is deeply flawed and breaks the premise with which this research was conceived. The same thing applies

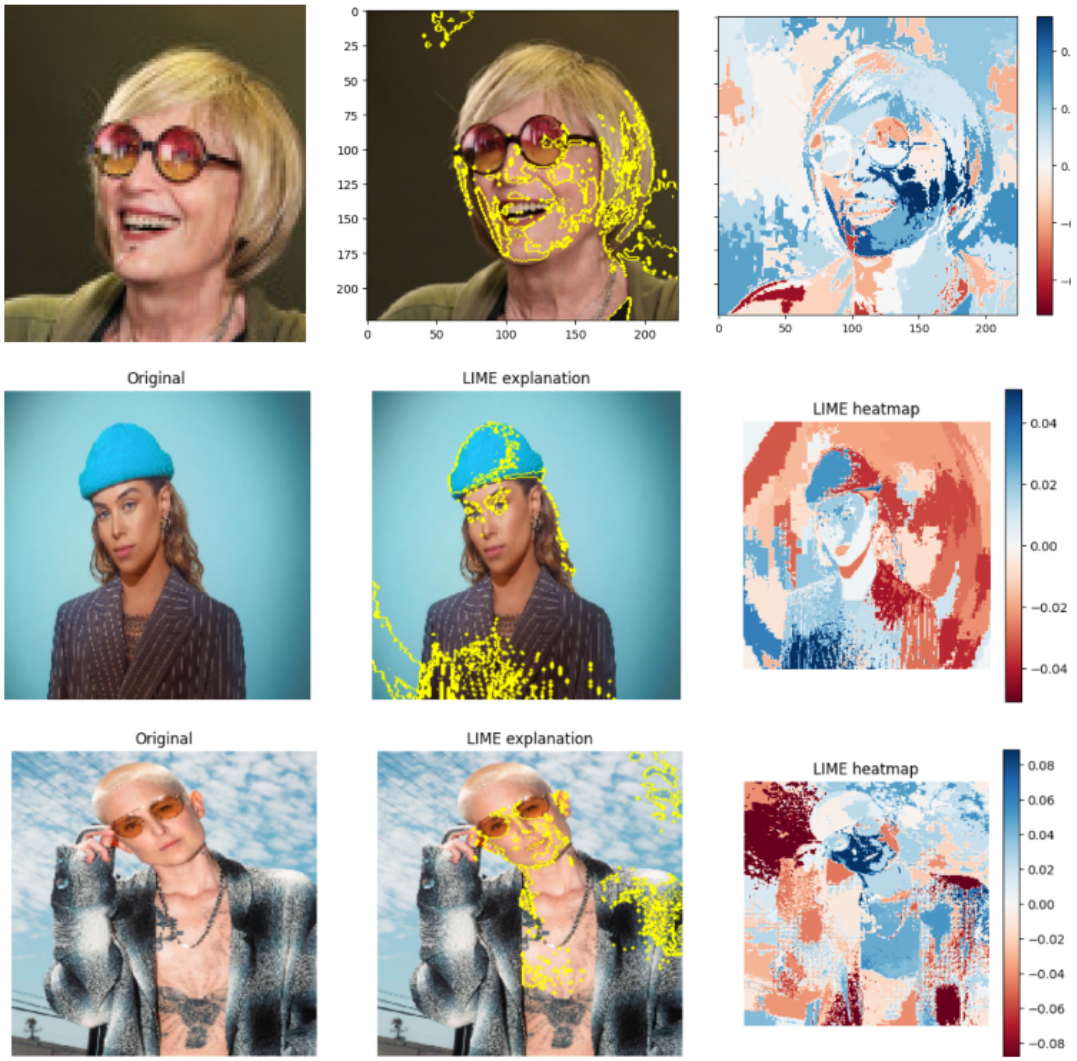


Figure 5.7: How clothing can impact the decision taken by the models

to the era to which an individual belongs, because the era defines the environment and the culture.

It must be said, however, that the discussion of gender as a performative act is very deep and complicated, and in this research its surface is scratched in order to motivate the work and observe whether the results obtained stick to what was initially thought.

It is necessary to ask: is the proposed work therefore ethical? Probably so, or at least, the intention with which the work was carried out is to create something that could actually make a contribution in changing a society that is hetero-centric and

where gender binarism is taken for granted. A society that neglects a significant portion of the population of which it is composed, that neglects, in doing so, the needs of people, the needs of individuals, by worsening their quality of life. Actual society struggles to give visibility to those who don't have it, swallows people and doesn't allow them to exist. The foundation of this work is based on this. To try to improve a technology that, despite being aprioristically unethical, exists, is used, and in its use discriminates against individuals.

5.3.1 Limitations

Although EfficientNet has shown promising results on our dataset, it is essential to shed light on the limitations of this approach and the general constraints of this technology when aiming for fairness and inclusivity in real-world applications.

The most important problem relates to data availability and diversity. The lack of data collected on individuals whose gender identity does not conform to a binary framework is a critical problem. In the dataset built for this research, despite the work done to expand it as best as possible, the low number of different gender non-conforming individuals does not allow a broad generalisation of the class, limiting the features that can lead to a correct classification in a real world scenario. Keeping in consideration the limitations of the dataset it is important to be aware of the consequences and of what the impact of such a technology might be on the societal level.

The Misgendering Machines, explained in section 2.3.3, underscores the potential pitfalls associated with these technologies. It is quite intuitive that attempting to determine a person's gender identity solely from an image is a complex task fraught with challenges and risks. The challenges stem from the intricate interplay of various factors including but not limited to facial features, cultural expressions, and individual self-perception. Moreover, gender identity itself is a nuanced and deeply personal aspect of human identity, often defying simplistic categorization or visual representation. Consequently, any algorithmic approach aimed at automatic gender recognition must navigate through a labyrinth of social constructs, biases, and ethical considerations to strive for inclusivity and fairness in its outcomes.

One major ethical concern revolves around the potential for misgendering individuals. Misgendering occurs when a machine incorrectly attributes a gender identity to a person based on facial features or other visual cues. This misclassification can be not only inaccurate but also deeply impactful for the individuals involved. Being misgendered by a machine has real-world consequences, as it can contribute to the perpetuation of gender stereotypes and reinforce biases that already exist in society [11].

Moreover the act of misgendering can have profound emotional and psychological repercussions on individuals, transcending mere misidentification to deeply impact

one's sense of self and belonging. When individuals are misgendered, whether intentionally or inadvertently, it can evoke feelings of frustration, invalidation, and alienation, exacerbating existing societal pressures and marginalization experienced by those whose gender identities diverge from societal binarism or expectations.

Furthermore, within the broader societal context, the misuse or inaccurate application of gender recognition technology carries significant implications for trust in automated systems and artificial intelligence. Persistent misgendering not only compromise the reliability and credibility of such systems but also perpetuate biases and reinforce dangerous stereotypes. This erosion of trust exacerbates disparities and inequalities in access to and utilization of technology across diverse communities, but also diminishes the potential benefits of technological advancements.

Therefore, addressing the ethical dimensions of gender recognition technology is not merely a technical challenge but a moral obligation, necessitating careful consideration of the societal impacts and ethical implications at every stage of development and implementation. Only through a concerted effort to prioritize inclusivity, fairness, and respect for individual autonomy is possible to mitigate the risks of misgendering and foster a more equitable technological landscape.

The ethical implications extend to the responsibility of technology users. It is crucial for individuals and organizations to be aware of the limitations and potential biases of gender inference technologies. Users should exercise caution and deploy such technologies only when absolutely necessary, ensuring that their use aligns with principles of fairness and respect for individuals' rights. Additionally, developers and policymakers need to prioritize the implementation of safeguards and regulations to mitigate the risks associated with misgendering and to promote the responsible use of gender inference technologies.

In conclusion, the ethical considerations surrounding the inference of gender identity from images highlight the need for a thoughtful and cautious approach. As technology continues to advance, it is imperative that developers, users, and policymakers collaborate to address these ethical concerns, foster transparency, and ensure that these technologies are deployed in a manner that respects the dignity and rights of individuals.

5.3.2 Future Works

The research presented explores a modern topic that presents both technical and ethical challenges. The lack of extended research in this specific topic and the recent increase in interest in the social issue proposed allows for a very high range of options when it comes to future works.

The first and simple one is the extension of the comparative analysis by shifting the focus on the evaluation of Vision Transformers [40]. Vision Transformers are a class of deep learning models that apply Transformer architecture [41],

originally designed for natural language processing, to computer vision tasks. With the introduction of self-attention mechanism it is possible to capture global dependencies between different section of the input image, unlocking a better contextual understanding. The advantages of Vision Transformers include their ability to capture long-range dependencies in images more effectively compared to traditional CNNs, as well as their flexibility in handling input of varying sizes through the use of the patch-based approach. The problem correlated with the usage of Vision Transformer in the task proposed in this research is that usually, in order to have a better understanding of the context, is necessary a large amount of training data, that, as explained before, is difficult to obtain.

Having achieved some interesting results, especially with EfficientNet, it is possible to think about potential uses for this system, keeping the ethical aspect in the first place.

One of the first idea is to use the proposed technology to evaluate the inclusivity of images used in official context. For instance, in a document created by a company it could be interesting to observe all the images in which people appears in order to give an inclusivity score of the document. This could be expanded, in the future, assuring a fair representation of minorities. This approach could be interesting in a world in which the communication is evolving to a more inclusive one, broadening the focus not only to the type of language and words used, but also to the choice of images that guarantee and are part of what is meant to be inclusive communication. In addition this approach would guarantee the anonymity of assignment to the individual, avoiding the problem of gender assignment on the individual, respecting the privacy of the individual itself and mitigating the inherent error in an algorithm of classification. The idea is to be as good as possible in the classification, but knowing the limitations intrinsic in a such complex task, mitigate the misgendering of individuals, contributing to a more inclusive communication.

A second idea is to use the proposed algorithm to obtain a broader statistic on the gender of individuals in a given society. An important lack of data lead to the problem of invisibility already explained in 2.2.1. "What gets counted counts" [42] so the idea to contribute to the visibility of gender non-conforming individual would fit optimally in this scenario. This could be done by, for example, creating a statistics using the images of social networks. The model would assign a gender to each individual, but the final output would be only the statistics, in order to keep the privacy of the single individual.

Another goal, but a little too far with the actual technology, would be to move to a broader perspective of gender recognition and to recognize gender as a continuous spectrum. The concept of the gender continuum, as discussed in "Gender Continuum" [43] recognizes that individuals everywhere can identify themselves in a diverse spectrum of gender expressions, transcending the limitations of the binary model. Such characteristics include gender identity, sexual orientation, scalable

personality traits such as assertiveness, inquisitiveness, empathy, and kindness, the observance of culturally defined behaviours such as dress, social interaction, and social roles and the physical traits associated with biological sex—including external genitalia, sex chromosomes, sex-related gene expression, sex hormones, and brain structure and activity—which vary along a continuum ranging from male to intersex to female, as well explained in Encyclopedia Britannica [44]. Categorical systems for recognizing gender are inherently unable to accurately represent the richness of gender diversity, risking misclassification and erasure of non-binary and gender-diverse individuals.

A potentially more ethical approach to the use of automatic gender recognition systems might be to classify people along a continuum that reflects the social spectrum between masculinity and femininity, rather than placing them in a few rigid categories. This approach could be of particular interest for purposes such as advertising, as it allows for a more nuanced and accurate representation of people's gender identities.

The main problem regarding this last approach is the technical difficulties associated. At the moment it is already very complicated to create a dataset that keep in consideration only three classes, so the creation of a more complete dataset appears to be utopian.

Chapter 6

Conclusions

The work proposed in this thesis is dedicated to the investigation into the intricacies of Automatic Gender Recognition system when trying to shift the perspective to a more ethical and fair use of these systems. The challenges faced were numerous, time-consuming, idea-consuming, and strenous in terms of the ethical questions that had to be asked.

The idea was to think a new paradigm regarding Automatic Gender Recognition systems, introducing a third class to represent gender non-conforming individuals, trying to improve the current state-of-the-art on this task. The objective was to modify the current Automatic Gender Recognition systems used everyday to make them more inclusive trying to make the need for an improvement in a technology coexist with the ethical problems inherent in assigning gender to a person based only on appearance. With the help of a variety of sociological studies and gender studies, it has been possible to learn more and more over the course of the last months, in order to respect the people involved in the study.

After the review of sociological and technical literature to gain awareness of the problem, the proposed work continued with the creation of an initial dataset. The dataset was not satisfactory, so it was expanded with particularly time consuming work.

The next step was fine-tuning 4 different computer vision models to assess accuracy on the proposed task. EfficientNet, with a 94% accuracy, was found to be the most accurate model, and consequently the analysis of the results from a qualitative point of view was carried out on the predictions of this model.

The analysis of the explanation obtained through LIME are a key part of the research and permit to observe whether the premises are met. Can the conception of gender as a performative act be used to accomplish Automatic Gender Recognition tasks? The analysis of the results explained further in 5.2 leads one to think that the premises are fulfilled. The question is when does the paradigm of gender as a performative act evolve rapidly and how much does it depend on the surrounding

environment. This evolution forces continuous adaptation to new paradigms in an attempt to maintain good performances.

Every care has been taken to adhere as best as possible to the ethical premises that guided this work trying to improve and subsequently rethink the use of technology that is difficult to justify in terms of its impact on the individual. Discussions were held with people from within the community, presenting the work to them to find critical issues, to explain the ideas written in the section 5.3.2, and to see if the efforts made were going in the right direction.

What has been achieved during these months has been satisfactory and by laying the groundwork for the ethical use of automatic gender recognition systems provides an opportunity to pursue particularly interesting future work.

Bibliography

[1] Surya Monro. «Non-binary and genderqueer: An overview of the field». In: *International Journal of Transgenderism* 20.2-3 (2019), pp. 126–131 (cit. on pp. 5, 8, 14).

[2] Kimberle Crenshaw. «"Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics"». In: *University of Chicago Legal Forum* (1989) (cit. on pp. 7, 13, 27).

[3] Jessica Ringrose Sara Bragg Emma Renold and Carolyn Jackson. «'More than boy, girl, male, female': exploring young people's views on gender diversity within and beyond school contexts». In: *Sex Education* 18.4 (2018), pp. 420–434 (cit. on p. 7).

[4] Anna Brown. *About 5% of young adults in the U.S. say their gender is different from their sex assigned at birth*. July 2022. URL: https://www.pewresearch.org/short-reads/2022/06/07/about-5-of-young-adults-in-the-u-s-say-their-gender-is-different-from-their-sex-assigned-at-birth (cit. on p. 7).

[5] UK Government. *LGBT Action Plan 2018: Improving the lives of Lesbian, Gay, Bisexual and Transgender people*. 2018. URL: https://www.gov.uk/government/publications/lgbt-action-plan-2018-improving-the-lives-of-lesbian-gay-bisexual-and-transgender-people/lgbt-action-plan-2018-improving-the-lives-of-lesbian-gay-bisexual-and-transgender-people (cit. on p. 8).

[6] Catherine D'Ignazio. *What gets counted counts*. 2020. URL: https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp/release/3 (cit. on pp. 8, 9).

[7] Jennifer A. Rode. «A theoretical agenda for feminist HCI». In: *Interacting with Computers* 23.5 (2011). Feminism and HCI: New Perspectives, pp. 393–400. ISSN: 0953-5438. DOI: https://doi.org/10.1016/j.intcom.2011.04.005. URL: https://www.sciencedirect.com/science/article/pii/S0953543811000385 (cit. on p. 9).

[8] E Matsuno. «Non-binary/Genderqueer Identities: a Critical Review of the Literature.» In: *Curr Sex Health Rep 9* (2017). DOI: https://doi.org/10.1007/s11930-017-0111-8 (cit. on p. 9).

[9] Educational Publishing Foundation. *Psychology of Sexual Orientation and Gender Diversity*. 2015. URL: https://journals.scholarsportal.info/browse/23290382/v02i0003 (cit. on p. 9).

[10] Eady A. Ross LE Dobinson C. «Perceived determinants of mental health for bisexual people: a qualitative examination.» In: *Am J Public Health.* (2010). DOI: 10.2105/AJPH.2008.156307 (cit. on p. 9).

[11] Os Keyes. «The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition». In: *Proceedings of the ACM on Human-Computer Interaction* 2 (Nov. 2018), pp. 1–22. DOI: 10.1145/3274357 (cit. on pp. 9, 55).

[12] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. «Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems». In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. , Montreal QC, Canada, Association for Computing Machinery, 2018, pp. 1–13. ISBN: 9781450356206. DOI: 10.1145/3173574.3173582. URL: https://doi.org/10.1145/3173574.3173582 (cit. on p. 9).

[13] Vedika Pareek. «Non-binary Gender and Data». In: (Dec. 2019). https://cis.pubpub.org/pub/binary-gender-data (cit. on p. 10).

[14] Patrick Farley. *What is the Azure AI Face service?* 2023. URL: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-identity (cit. on p. 11).

[15] Sarah Bird. *Responsible AI investments and safeguards for facial recognition*. 2022. URL: https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/ (cit. on p. 12).

[16] S. De Beauvoir, C. Borde, and S. Malovany-Chevallier. *The Second Sex*. Knopf Doubleday Publishing Group, 2011. ISBN: 9780307277787. URL: https://books.google.it/books?id=_hywlrNuYvIC (cit. on pp. 12, 53).

[17] Candace West and Don H. Zimmerman. «Doing gender». In: *Gender Society* 1.2 (1987), pp. 125–151. DOI: 10.1177/0891243287001002002. URL: http://journals.sagepub.com/doi/10.1177/0891243287001002002 (cit. on pp. 13, 53).

[18] Judith Butler. «Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory». In: *Theatre Journal* 40.4 (1988), pp. 519–531. ISSN: 01922882, 1086332X. URL: http://www.jstor.org/stable/3207893 (visited on 10/22/2023) (cit. on pp. 14, 53).

[19] Wikipedia. *Performative utterance*. URL: https://en.wikipedia.org/wiki/Performative_utterance#:~:text=Austin's%20definition,-In%20order%20to&text=Performative%20utterances%20are%20not%20true,wrong%20they%20are%20%22happy%22. (cit. on p. 14).

[20] Ambika. *What is Computer Vision? (History, Applications, Challenges)*. URL: https://medium.com/@ambika199820/what-is-computer-vision-history-applications-challenges-13f5759b48a5 (cit. on p. 18).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep Residual Learning for Image Recognition». In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385 (cit. on pp. 19, 20, 38, 46).

[22] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. «MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications». In: *CoRR* abs/1704.04861 (2017). arXiv: 1704.04861. URL: http://arxiv.org/abs/1704.04861 (cit. on pp. 19, 23, 24, 38, 46).

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. *Going Deeper with Convolutions*. 2014. arXiv: 1409.4842 [cs.CV] (cit. on pp. 19, 21–23, 38).

[24] Mingxing Tan and Quoc Le. «EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks». In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html (cit. on pp. 19, 25, 38).

[25] Y. Bengio, P. Simard, and P. Frasconi. «Learning long-term dependencies with gradient descent is difficult». In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166. DOI: 10.1109/72.279181 (cit. on p. 19).

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. «Rethinking the Inception Architecture for Computer Vision». In: *CoRR* abs/1512.00567 (2015). arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567 (cit. on p. 22).

[27] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O'Toole. «An other-race effect for face recognition algorithms». In: *ACM Trans. Appl. Percept.* 8.2 (Feb. 2011). ISSN: 1544-3558. DOI: 10.1145/1870076.1870082. URL: https://doi.org/10.1145/1870076.1870082 (cit. on p. 26).

[28] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. *Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings*. 2018. arXiv: 1809.02169 [cs.CV] (cit. on p. 26).

[29] Joy Buolamwini and Timnit Gebru. «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification». In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 23–24 Feb 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html (cit. on p. 26).

[30] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. «On the Robustness of Face Recognition Algorithms Against Attacks and Bias». In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (Apr. 2020), pp. 13583–13589. DOI: 10.1609/aaai.v34i09.7085 (cit. on p. 26).

[31] Jonathan Frankle Clare Garvie Alvaro Bedoya. *The Perpetual Lineup*. URL: https://www.perpetuallineup.org/ (cit. on p. 26).

[32] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. «Face Recognition Performance: Role of Demographic Information». In: *IEEE Transactions on Information Forensics and Security* 7.6 (2012), pp. 1789–1801. DOI: 10.1109/TIFS.2012.2214212 (cit. on pp. 27, 28).

[33] Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. «Gender Classification and Bias Mitigation in Facial Images». In: *CoRR* abs/2007.06141 (2020). arXiv: 2007.06141. URL: https://arxiv.org/abs/2007.06141 (cit. on pp. 27, 29, 32–34).

[34] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. «"Why Should I Trust You?": Explaining the Predictions of Any Classifier». In: *CoRR* abs/1602.04938 (2016). arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938 (cit. on pp. 31, 36, 41, 42, 47).

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. «Deep Learning Face Attributes in the Wild». In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015 (cit. on pp. 31, 32).

[36] Wikipedia. *List of people with non-binary gender identities*. URL: https://en.wikipedia.org/wiki/List_of_people_with_non-binary_gender_identities (cit. on pp. 32, 34).

[37] Damien Dablain, Kristen N. Jacobson, Colin Bellinger, Mark Roberts, and Nitesh Chawla. *Understanding CNN Fragility When Learning With Imbalanced Data*. 2022. arXiv: 2210.09465 [cs.CV] (cit. on p. 32).

[38] M. Stone. «Cross-Validatory Choice and Assessment of Statistical Predictions». In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), pp. 111–147. ISSN: 00359246. URL: http://www.jstor.org/stable/2984809 (visited on 02/20/2024) (cit. on p. 38).

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. «ImageNet: A large-scale hierarchical image database». In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on p. 38).

[40] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV] (cit. on p. 56).

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL] (cit. on p. 56).

[42] Lauren Klein Chaterine D'Ignazio. *Data-feminism mitpress*. URL: https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp/release/3 (cit. on p. 57).

[43] Selin Gülgöz, Deja L. Edwards, and Kristina R. Olson. «Between a boy and a girl: Measuring gender identity on a continuum». English (US). In: *Social Development* 31.3 (Aug. 2022), pp. 916–929. ISSN: 0961-205X. DOI: 10.1111/sode.12587 (cit. on p. 57).

[44] Duignan Brian. *gender continuum*. Oct. 2023. URL: https://www.britannica.com/topic/gender-continuum (cit. on p. 58).