BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

THE IMPACT OF NEURAL NETWORK VISUALIZATIONS ON USER DECISION MAKING IN DEEPFAKE DETECTION USING HEATMAPS

VLIV VIZUALIZACÍ NEURONOVÝCH SÍTÍ

NA ROZHODOVÁNÍ UŽIVATELŮ PŘI DETEKCI DEEPFAKE POMOCÍ HEATMAP

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. FILIP BRNA**
AUTOR PRÁCE

SUPERVISOR **Ing. MILAN ŠALKO**
VEDOUCÍ PRÁCE

BRNO 2025

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

164756

Institut:	Department of Intelligent Systems (DITS)
Student:	**Brna Filip, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Cybersecurity
Title:	**The impact of neural network visualizations on user decision making in deepfake detection using heatmaps**
Category:	Security
Academic year:	2024/25

Assignment:

1. Familiarise yourself with deepfake content detection techniques, particularly with regard to facial recognition. Investigate research on methods for the explainability of these techniques (focus on methods utilizing heatmaps).
2. Study the problem of decision support systems (DSS) in the area of facial deepfakes detection.
3. Design an experiment in which people make assisted decisions using heatmaps from a deepfake detector. Do the experiment on 3 groups (control group, group with output from the detector, group with heatmaps and output from the detector). The minimum size of each group must be at least 15 people. Focus on how the use of heatmaps affects the user's decision-making processes. Select two detectors from which a heatmap of explainability can be obtained.
4. Train two selected neural network models to detect deepfakes and create heatmaps that describe the detector's decision making. Prepare a set of at least 30 samples for each detector to be used in the experiment.
5. Conduct the proposed experiment.
6. Evaluate the results obtained from the experiment and discuss the outcomes.

Literature:

- Visual Analytics for Explainable Deep Learning, Jaegul Choo, Shixia Liu https://doi.org/10.48550/arXiv.1804.02527
- Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection, Kiran Raja, Ying Xu DOI:10.1109/WACVW54805.2022.00044
- Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms, P. Jonathon Phillips https://doi.org/10.1073/pnas.172135511

Requirements for the semestral defence:
Items 1 to 3

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Šalko Milan, Ing.**
Head of Department:	Kočí Radek, Ing., Ph.D.
Beginning of work:	1.11.2024
Submission deadline:	21.5.2025
Approval date:	8.4.2025

Abstract

Deepfake technology is a growing threat, with AI detection tools often lacking transparency. This paper examines the impact of XAI, specifically heatmaps (Grad-CAM++), on human decision-making in deepfake face detection. The aim was to verify how AI scores and heatmaps affect user accuracy and confidence, especially when AI fails.

A three-group experiment (control, AI score, AI score + heatmap) showed a 67% baseline human success rate (F1). AI assistance did not significantly improve F1 overall. The detector score strongly influenced users, leading to worse results when AI errors occurred. The key finding is that adding a heatmap mitigated this negative impact, maintained control-level accuracy when AI errors occurred, improved confidence-weighted scores, and led to more balanced decision-making. Despite the objective benefits, users subjectively rated heatmaps as of little use.

The work demonstrates heatmaps' potential to improve human decision-making robustness with AI support. It also highlights the need to develop more user-friendly XAI methods to combat deepfakes effectively.

Abstrakt

Deepfake technológia je rastúcou hrozbou, pričom AI detekčné nástroje často postrádajú transparentnosť. Táto práca skúma vplyv XAI, konkrétne heatmap (Grad-CAM++), na ľudské rozhodovanie pri detekcii deepfake tvárí. Cieľom bolo overiť, ako AI skóre a heatmapy ovplyvňujú presnosť a dôveru používateľov, najmä pri zlyhaní AI.

Experiment s tromi skupinami (kontrolná, AI skóre, AI skóre + heatmapa) ukázal 67% základnú ľudskú úspešnosť (F1). AI asistencia celkovo F1 významne nezlepšila. Skóre detektora silne ovplyvňovalo používateľov a pri chybách AI viedlo k horším výsledkom. Kľúčovým zistením je, že pridanie heatmapy mitigovalo tento negatívny dopad, udržalo presnosť na úrovni kontrolnej skupiny pri chybách AI, zlepšilo skóre vážené istotou a viedlo k vyváženejšiemu rozhodovaniu. Napriek objektívnym prínosom, používatelia subjektívne hodnotili heatmapy ako málo užitočné.

Práca demonštruje potenciál heatmap zlepšiť robustnosť ľudského rozhodovania s podporou AI. Zároveň poukazuje na potrebu vývoja používateľsky zrozumiteľnejších XAI metód pre efektívny boj proti deepfake.

Keywords

Deepfake Detection, Explainable AI, XAI, Heatmaps, Grad-CAM, Human-AI Interaction, User Study, Decision Making, Visualization

Klíčová slova

Detekcia deepfake, Vysvetliteľná umelá inteligencia, XAI, Teplotné mapy, Grad-CAM, Interakcia človek-AI, Užívateľská štúdia, Rozhodovanie, Vizualizácia

Reference

BRNA, Filip. *The impact of neural network visualizations on user decision making in deepfake detection using heatmaps.* Brno, 2025. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Milan Šalko

Rozšířený abstrakt

Technológia deepfake, umožňujúca vytváranie hyperrealistických syntetických médií, predstavuje dvojsečnú zbraň. Na jednej strane ponúka nové možnosti v kreatívnom priemysle či vzdelávaní, na strane druhej prináša vážne hrozby v podobe šírenia dezinformácií, krádeží identity a podkopávania dôvery v digitálny obsah. S rastúcou sofistikovanosťou metód na vytváranie deepfake médií je nevyhnutné vyvíjať účinné detekčné nástroje. Systémy založené na hlbokých neurónových sieťach dosahujú sľubné výsledky, avšak ich fungovanie predstavuje "čierne skrinky" bez dostatočnej transparentnosti, čo vedie k nízkej vysvetliteľnosti, obmedzenej dôvere používateľov a pomalšej adopcii. Vysvetliteľná umelá inteligencia (XAI) ponúka riešenia na zvýšenie transparentnosti a dôveryhodnosti týchto systémov.

Táto diplomová práca sa zameriava na preskúmanie úlohy a vplyvu XAI mechanizmov, konkrétne vizualizácií vo forme heatmap generovaných metódou Grad-CAM++, na proces ľudského rozhodovania pri detekcii deepfake obrázkov tvárí. Hlavným cieľom bolo navrhnúť, realizovať a vyhodnotiť experiment, ktorý by odpovedal na nasledujúce výskumné otázky: (RQ1) Do akej miery dokážu ľudia sami rozlíšiť pravé a falošné obrázky? (RQ2) Aký je vplyv poskytnutia výstupného skóre AI detektora a Grad-CAM++ heatmapy na presnosť, istotu a rozhodovací proces používateľov? Práca testovala hypotézy predpokladajúce negatívny dopad absencie pomoci (H1), silný vplyv výstupu detektora na rozhodovanie (H2) a schopnosť heatmap zmierniť slepú dôveru v AI a podporiť tým informovanejšie rozhodnutia (H3).

Prostredníctvom online platformy bol realizovaný experiment formou používateľskej štúdie s 204 platnými účastníkmi, náhodne rozdelenými do troch skupín: 1. *Kontrolná skupina* (bez asistencie), 2. *Detektor skupina* (poskytnuté len percentuálne skóre AI detektora udávajúce pravdepodobnosť deepfake), 3. *Det. & Heatmap skupina* (poskytnuté skóre aj Grad-CAM++ heatmapa vizualizujúca oblasti záujmu detektora). Ako detektory boli použité modely EfficientNet-B3 a B4 natrénované na datasete FaceForensics++. Kľúčovým prvkom dizajnu bol výber 60 testovacích obrázkov (30 pravých, 30 falošných), ktoré boli strategicky zvolené tak, aby rovnomerne pokrývali celé spektrum spoľahlivosti detektora (0-100%) a zahŕňali aj náročné "šedé zóny" – prípady nízkej istoty AI alebo jej chybnej klasifikácie. Úlohou účastníkov bolo pre každý obrázok rozhodnúť, či je pravý alebo falošný, a ohodnotiť svoju istotu na 5-bodovej Likertovej škále. Okrem štandardných metrík (presnosť, F1-skóre, recall, specificity, precision) bola použitá aj špecifická metrika Confidence-Weighted Score (CWS), ktorá zohľadňuje správnosť aj istotu odpovede. Boli tiež zbierané demografické údaje, informácie o predchádzajúcich skúsenostiach s tvorbou/detekciou deepfakes a subjektívna spätná väzba po experimente.

Analýza výsledkov ukázala, že základná ľudská schopnosť rozlíšiť deepfake (RQ1) bola na úrovni približne 67% (F1-skóre), čo je síce významne lepšie ako náhodné hádanie, avšak sprevádzané štatisticky významným sklonom častejšie klasifikovať obrázky ako falošné. Hypotéza H1 bola zamietnutá, keďže v celkovom F1-skóre neboli medzi skupinami nájdené štatisticky významné rozdiely v tomto experimentálnom nastavení. AI asistencia sa však ukázala ako prospešná v prípadoch, keď bola predikcia AI správna. Hypotéza H2 bola potvrdená – výstupné skóre AI detektora malo silný a merateľný vplyv na rozhodovanie účastníkov zaradených do *Detektor* a *Det. & Heatmap skupiny*, títo účastníci mali tendenciu nasledovať odporúčanie AI, najmä pri vysokej úrovni jej istoty. Tento vplyv mal však aj negatívne dôsledky: pri nesprávnych predikciách AI dosiahla *Detektor skupina* (len skóre) štatisticky významne horšiu presnosť a CWS ako *Kontrolná skupina*, čo potvrdzuje riziko slepého nasledovania zavádzajúcich informácií. Hypotéza H3 bola silne podporená. Pri-

danie Grad-CAM++ heatmapy dokázalo efektívne mitigovať negatívny dopad nesprávneho skóre AI. Presnosť *Det. & Heatmap skupiny* pri chybách AI nebola štatisticky horšia ako *Kontrolná skupina* (na rozdiel od *Detektor skupiny*) a CWS bolo v týchto prípadoch štatisticky významne lepšie v *Det. & Heatmap skupine* ako v *Detektor skupine*. *Det. & Heatmap skupina* bola tiež jediná, ktorá nevykazovala štatisticky významnú nerovnováhu medzi metrikami recall, specificity a precision, čo naznačuje informovanejší a vyváženejší rozhodovací proces. Zaujímavým paradoxom však bolo, že napriek objektívnym prínosom účastníci tejto skupiny subjektívne hodnotili heatmapy ako málo užitočné a zároveň uvádzali štatisticky významne vyššiu mieru spoliehania sa na AI systém (skóre + heatmapa) ako *Detektor skupina*.

Táto práca poskytuje empirické dôkazy o komplexnej interakcii medzi ľuďmi a AI systémami pri detekcii deepfake. Potvrdzuje sa, že zatiaľ čo AI asistencia môže byť prospešná, poskytovanie samotného skóre spoľahlivosti bez ďalšieho kontextu je rizikové, najmä v prípadoch chýb AI. Vysvetliteľnosť vo forme Grad-CAM++ heatmap sa ukázala ako účinný nástroj na mitigáciu týchto rizík, zlepšenie kvality rozhodovania a podporu vyváženejšieho prístupu. Nesúlad medzi objektívnymi výsledkami a subjektívnym vnímaním užitočnosti heatmap však poukazuje na kritickú výzvu pre oblasť XAI – potrebu navrhovať nielen technicky funkčné, ale predovšetkým používateľsky zrozumiteľné a intuitívne metódy vysvetlenia. Práca prispieva k lepšiemu pochopeniu dynamiky človeka a AI v procese rozhodovania a poskytuje východiská pre ďalší výskum a vývoj robustnejších a dôveryhodnejších detekčných systémov.

The impact of neural network visualizations on user decision making in deepfake detection using heatmaps

Declaration

I hereby declare that this Master's thesis was prepared as an original work by the author under the supervision of Ing. Milan Šalko. The supplementary information was provided by Ing. Anton Firc. I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis. Assistance with grammar and style checking was provided by the Grammarly tool during the writing process.

.....
Filip Brna
May 17, 2025

Acknowledgements

My sincere thanks go to my supervisor, Ing. Milan Šalko, for his expert guidance and support.

I am deeply grateful to my parents, sister, and whole family for their unwavering love and encouragement.

A BIG thank you to my girlfriend for her patience and support throughout my studies.

I also appreciate my friends, especially those who participated in the research experiment.

Computational resources were provided by MetaCentrum.

Contents

List of Figures

Chapter 1

Introduction

The exponential growth of artificial intelligence and machine learning technologies has brought significant innovations across various fields. One of the areas is the emergence of deepfake technology, which has a significant impact and consequences for the creative industry and cybersecurity. Deepfakes are hyperrealistic synthetic media created through technologies using generative adversarial networks and diffusion models. Although their use offers potential benefits in areas such as entertainment and education, their misuse can pose serious threats, including spreading disinformation, identity theft, or compromising security systems.

The techniques for creating deepfakes are increasingly sophisticated, and it is therefore important to respond accordingly by constantly researching methods for their detection. Automated detection systems using advanced neural network architectures have shown considerable success in this area. However, these systems often function as a black-box model, where any transparency of their decision-making processes is absent. Lack of explainability can lead to a loss of trust in detection systems from users and thus slow down their wider adoption, especially by non-experts.

Explainability mechanisms like heatmaps have been integrated into detection systems to address this challenge. Heatmaps help to highlight the parts of the image that had the most significant impact on the final decision of the detection models, thereby increasing their transparency and providing users with valuable insights. This thesis investigates the role of explainability mechanisms (specifically Grad-CAM++ heatmaps) and their impact on human decision-making processes in deepfake detection.

The research objectives include designing and implementing an experiment involving three groups: a control group, a group with the output from the deepfake detector at its disposal, and a group that will be assisted in its decision-making by heatmaps and the outputs from the deepfake detector. Crucially, the experiment was designed to include challenging "gray zone" cases where the AI detector was uncertain or even incorrect, as these situations are critical for the human classification process. The experiment aims to evaluate to what extent the detector score and heatmaps explanation can influence the ability of participants to distinguish between genuine and fake facial images and how they affect reliance on AI. Preliminary findings suggest a complex relationship: while AI assistance can be beneficial when the detector is correct, relying solely on AI scores carries risks when the AI errs. Explainability through heatmaps plays a crucial role in mitigating these risks and fostering more informed, balanced decisions, although their perceived usefulness by users presents an interesting paradox. The summarization of the effectiveness of heatmaps in

supporting human decisions in this work can also contribute to developing more trustworthy and user-friendly deepfake detection systems.

The work begins with a theoretical foundation, which includes the technologies and techniques used to create facial deepfakes. Next, the work describes the methods used for manual and automatic deepfakes detection, the importance of explainability methods in the detection process, examples of technologies used to create heatmaps, and the challenges related to explainability. The work continues with defining a decision support system and its role and challenges in the human-AI decision-making process. The thesis then moves on to the experimental part, which includes the experimental design, implementation, and evaluation of the results.

The findings provide valuable insights into the dynamics of user decision-making when interacting with imperfect AI detectors and specifically assess the effectiveness and user perception of Grad-CAM++ heatmaps as an explainability tool in this context, offering recommendations for future development of AI-assisted deepfake detection tools.

Chapter 2

Deepfakes

The name combines two words: deep learning[1] and fake. The phenomenon of deepfake itself offers new possibilities for creative and productive use in various industries, but it is not without threats in cyberspace and other unethical aspects [60]. It is credible media, visual or audio, generated mainly by a deep neural network, which aims to be realistic in the eyes of humans. This medium reproduces fictional events by replacing another person's voice or face in a picture or video. Deepfakes can be broadly categorized into two main domains: the **Facial domain** and the **Speech domain**, as shown in Figure 2.1.

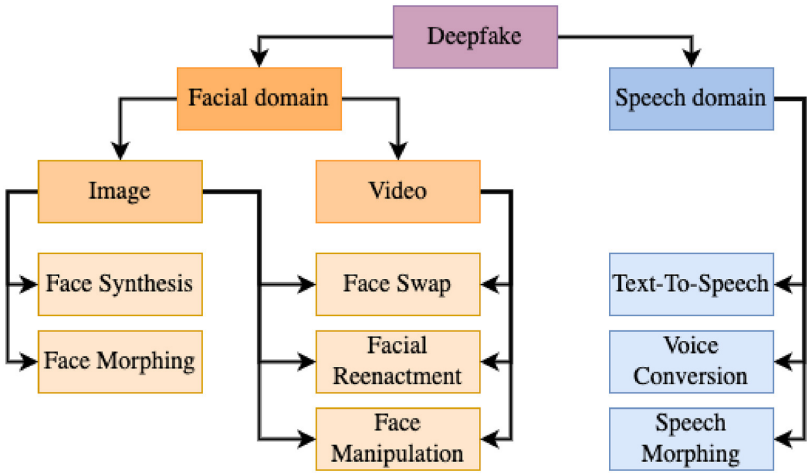


Figure 2.1: The deepfakes are categorized into facial and speech domains, further broken down into subcategories that illustrate individual approaches. The figure was taken from [41].

Most facial deepfakes are created by replacing a face with a fake image of someone else. Still, another approach can also be used to generate the face of a nonexistent person with the help of artificial intelligence (AI)[2] and machine learning (ML)[3]. In today's disin-

[1] Deep learning allows computational models composed of multiple processing layers to learn data representations with multiple levels of abstraction. [65]

[2] Artificial intelligence enables computers and machines to simulate human learning, comprehension, problem-solving, decision-making, creativity, and autonomy. [100]

[3] Machine learning is a branch of artificial intelligence focused on enabling computers and machines to imitate how humans learn, perform tasks autonomously, and improve their performance and accuracy through experience and exposure to more data. [54]

formation age, the deepfake is often used to polarise society, politically or otherwise. In addition to deceiving people, deepfakes can also be used to deceive the security features of various security systems, which can lead to financial damage or loss of assets or valuable information. This was a sign that the popularity of deepfake has started to grow even in academia. At the same time, in 2017, only a few dozen scientific papers were published; nowadays, hundreds to thousands of publications per year can be seen [74]. Figure 2.2 illustrates the trend in search interest for the term "deepfake," showing that awareness of the term was zero until 2017 but gradually increased until 2023, when it peaked. Since then, popularity has remained relatively high, consistent with advances in generative AI and increasing societal awareness.

Figure 2.2: Google Trends search interest for the term "deepfake" (Jan 2017 – Dec 2024). The numbers represent search interest relative to the highest point, where 100 is the peak popularity.

In this chapter, the facial deepfakes will be discussed further.

2.1 Technologies

Deep generative models are behind the revolution in manipulating and generating fake images. The generation process uses advanced machine learning architectures, which, due to their efficiency, cause problems in determining the boundary between the genuine people in the image and the fake ones. In the generation domain, there are two main approaches, Generative Adversarial Networks (GANs) and Diffusion Models, which are prominent among them. The above approaches have their strengths and weaknesses, and this subsection will describe both and how they contribute to the generation of deepfake images.

2.1.1 Generative Adversarial Network

A generative adversarial network is a type of neural network[4] that attempts to generate synthetic images based on training data. GANs are used, for example, in face synthesis or face swapping, where this approach uses semi-supervised and unsupervised learning

[4]A neural network is a machine learning program or model that makes decisions like the human brain by using processes that mimic how biological neurons work together to identify phenomena, weigh options, and arrive at conclusions. [1]

techniques and thus provides the possibility of training without extensive annotation of the training data.

This approach can be characterized as a pair of networks, a generator, and a discriminator, typically implemented by multilayer networks consisting of convolution and/or fully connected layers. These networks are called competing networks. The generator does not have access to the actual images during training, only to the interaction with the discriminator, which is why it produces ever-new, more authentic fakes. Meanwhile, the discriminator is trained to discriminate between originals and generator-generated fakes and can access synthetic and natural samples. Thus, both mentioned networks are trained simultaneously. In a basic GAN, the discriminant network can similarly be characterized as a function that maps the input image data to the probability (interval 0-1, where 0 is fake and 1 is genuine) of whether the image is genuine or not. Based on the evaluation result, this probability is then used to train the generator, leading it to improve its ability to generate fakes of better quality. The generator creates forgeries in order to create images that are very similar to the genuine ones and send them to a discriminator who tries to evaluate them. [27] A diagram of the process is shown in Figure 2.3.

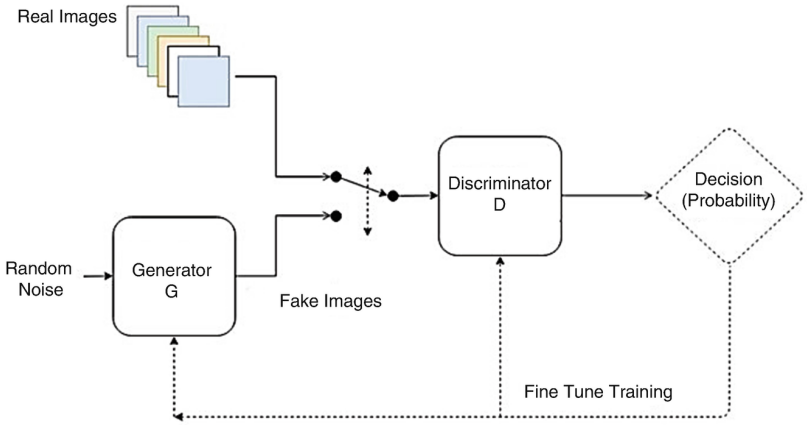


Figure 2.3: General block diagram of a generative adversarial network. The figure was taken from [87].

There are several GAN manipulation approaches, and each of them excels in a different domain of manipulation; these are AttGAN, StyleGAN, and STGAN. AttGAN allows detailed attribute-based manipulation while preserving identity, making it suitable for age manipulation, for example. StyleGAN, on the other hand, allows independent and precise changes to be made to different facial features by using a feature called style blending, where attributes can be applied at different layers of the network. Higher layers deal with broader cue attributes, while lower layers focus on finer details. This layered approach allows face manipulation with high fidelity. STGAN provides selective feature modification while leaving other aspects of the face unchanged; the disadvantage is that more features are selected for modification, and the model may exhibit poorer quality. [75, 86]

GANs, together with variational autoencoders (VAEs)[5], also play an important role in the reenactment. GAN modifies or generates a medium with realistic facial and body movements, and its discriminator is used to improve the quality. This model of generating

[5]Variational autoencoders are generative models used in machine learning to generate new data in the form of variations of the input data they're trained on. In addition to this, they also perform tasks common to other autoencoders, such as denoising. [100]

and then assessing the quality measure allows for a gradual and significant increase in quality during reconstruction using deep learning.

2.1.2 Diffusion model

Diffusion models belong to the family of deep generative models and are considered a new state-of-the-art technology. In recent years, research in the field of diffusion has grown considerably, and nowadays, even diffusion models have taken over from the long-dominant GANs. These models represent a significant advance because they allow more realistic and efficient options for generation and also do not need to train additional discriminators like GAN or align the posterior distribution like VAE [23]. Models of this type can create hyperrealistic facial deepfakes that are unrecognizable to the naked eye. They can eliminate defects such as imperfect edges of the face and blurring and also correct the problem with the symmetry of the characteristic features of the face. [10, 24] The principle of diffusion models uses two Markov chains[6] [117]:

1. **Forward chain:** Takes care of the gradual distortion of the data by adding noise. This chain is usually designed by hand with the task of transforming any data distribution into a simple distribution.

2. **Reverse chain:** This chain inverts the original distribution by learning transition kernels parameterized by deep neural networks. This uses Denoising Diffusion Probabilistic Models [52], which takes care of the opposite process and thus converts noise back into data.

New data points are generated by reconstructing the original data distribution, starting from noise and iteratively refining it through the reverse diffusion process. The entire process of the diffusion model is shown in Figure 2.4.

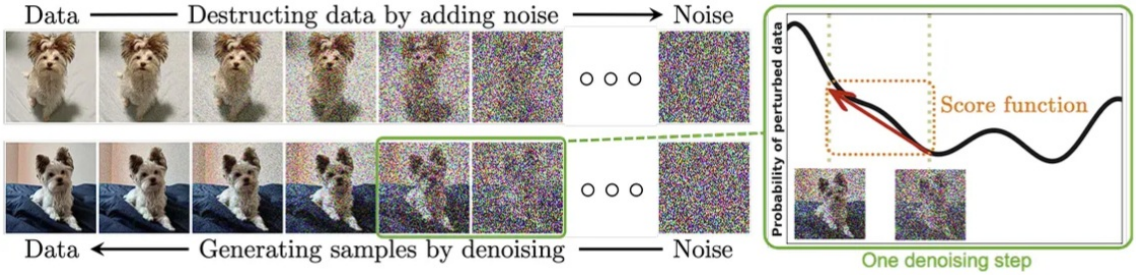


Figure 2.4: Diffusion models distort the input image by gradually adding noise, and then this process is reversed to generate new data from the noise. Each step of backward denoising typically requires estimating a score function, a gradient pointing to data directions with higher probability and less noise. The figure was taken from [117].

The process can also be improved using the Latent Diffusion Model [88], which improves the process by generating vectors in the latent space to speed up the diffusion process and handle complex data distributions.

Diffusion-generated images are categorized into two groups [66]:

1. **Text-guided image editing or generation:** Modifies an already existing image based on the entered text, e.g., InstructPix2Pix [17], Imagic [59].

[6] Markov chain is a process where the next state depends only on the current state.

2. **Text-to-image:** Generates an image based on the entered text containing detailed specifications of the resulting media, e.g., DALL-E [77], Imagen [51].

2.2 Face Deepfakes

This section will explain the techniques used to create face deepfakes. Four of the most used techniques: reenactment, synthesis, substitution, and manipulation, are briefly described in Figure 2.5. Each approach will be described with tools, usable datasets[7], detection methods, usage, and attack threats. [41]

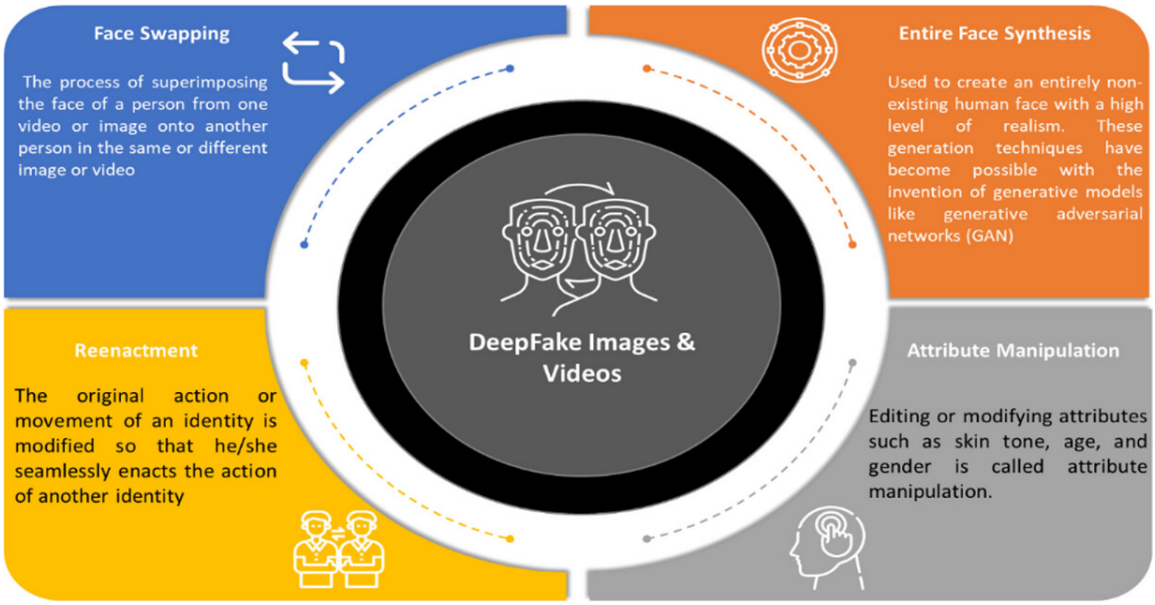


Figure 2.5: Various forms of image and video deepfake creation techniques. The figure was taken from [33].

2.2.1 Facial reenactment

This approach is one of the most complex and exciting applications for generating deepfakes, given that its essence is to change the complex details of expressions or movements, which must appear natural; otherwise, even minor inaccuracies can lead to detection.

The key concepts are the source identity, which provides the expressions/actions, and the target identity, which is the person whose visual representation changes based on the source identity.

The result is a realistic transfer of expressions corresponding to the source identity to the target identity, which retains its baseline identity. This is in contrast to the face-swap approach, where the face of the source replaces the face of the target. Depending on the degree of reenactment, facial reenactment can focus on transferring facial expressions, head movements, and eye movements, or facial reenactment, which is about capturing and using the body posture and movements of the source identity. This target identity can be synchronized with any accompanying sound, such as video. In reconstruction, it is crucial

[7]The dataset contains the data from which the AI learns, which is just as, if not more, important than the AI algorithms themselves.

to correctly locate the essential points such as eyes, nose, mouth, joints. [33, 41] An example of reenactment is shown in Figure 2.6.

Figure 2.6: The source identity provides a facial expression, which is then projected onto the target identity, represented by actor Daniel Craig, using a reenactment. The figure was taken from [107].

Not to forget the contraction of the facial muscles such as the inner brow raiser, upper brow raiser, and jaw drop, each set for very short moments. Facial Action Coding System has over 7000 combinations, and it is necessary to transform them into numeric data, which are later used to improve the quality of the reenactment results. [33, 41]

Tools and datasets

It is possible to observe a smaller number of reenactments due to the lower popularity of this approach compared to the face swap. All tools are listed in Table 2.1, which gives an overview of tools using different approaches to create deepfakes using reenactment. These tools exist as open-source solutions without additional user interfaces, which puts much higher demands on potential users. In Table 2.1 in the datasets section, it is possible to see a selection of datasets containing mostly videos on which various human face animations are captured. [41]

Usage and Attack vectors

Like all others, this approach has practical uses but poses a serious security risk. It allows for identity theft, reputation damage, and fraud using highly realistic videos accompanied by speech created from an image or impersonating other people in real-time, for example, during video conferences or job interviews, as the FBI warned in 2022 [7]. However, there are many more real-world examples; an attempt to reenact a deepfake financial attack was recorded on the chief executive officer of a company with a website selling fitness supplements [26], and other examples served as preventive measures to familiarize people with what deepfake technology is capable of, deepfake Queen Elizabeth II. [19] or former US President Barack Obama [105]. Deepfakes thus compromise security systems and features that rely on live interaction, such as Know Your Customer (KYC). Legitimate uses include virtual reality and education, which transfer information more reliably and immersively

Table 2.1: The table shows facial reenactment tools with their publications and datasets. The table was taken from [41] and modified.

Tools	
Name	**Link**
StyleMask [16]	https://github.com/StelaBou/StyleMask
FDGLS [15]	https://github.com/StelaBou/stylegan_directions_face_reenactment
AVFR [3]	http://cvit.iiit.ac.in/research/projects/cvit-projects/avfr
Face2Face [107]	https://github.com/datitran/face2face-demo
ATVGnet [22]	https://github.com/lelechen63/ATVGnet
Datasets	
Name	**Content**
DeepFake MNIST+	10,000 facial animation videos in ten different actions
FaceForensics	500,000 edited images
FaceForensics++	1000 original video sequences that have been manipulated with: Deepfakes, Face2Face, FaceSwap and NeuralTextures

and allow simulating real-life scenarios like a video with David Beckham [30] that spread awareness about malaria in 9 foreign languages. [33]

Detection

The detection of deepfakes is primarily based on the identification of unique biological and also behavioral patterns in the created medium; these patterns are then evaluated. Therefore, analyzed deepfakes can show subtler and even more significant inconsistencies and artifacts, subsequently used for detection using the selected method. Individual detection models, such as multi-stream networks, learn regional artifacts unique to the given type of manipulation. In addition to artifacts, biological signals are also used for detection. Eye movements, blink frequency, head position, and emotions are analyzed, and the aim is to detect signs of unnatural human behavior. Another fundamental detection approach can be the analysis of audiovisual synchronization, i.e., whether the movement of the mouth and spoken phonemes are coordinated. Combining all the methods mentioned above makes it possible to better distinguish genuine media from deepfakes. [33, 41]

2.2.2 Face synthesis

The Synthesis-oriented approach generates faces of nonexistent persons based on learned attributes and inputs. These realistic deepfakes are created using Diffusion models, Generative Adversarial Networks (GANs), or hybrid GANs, which allow parameterization and thus influence the appearance of the final media. Parameterization is not always necessary;

an example of such a parameterless synthesis can be seen in Figure 2.7, where it is possible to see how high a level of verisimilitude deepfakes with this approach achieve. [34, 56]

Figure 2.7: The image shows six faces created using the website thispersondoesnotexist.com, synthesizing a non-existent human face with each page refreshes. a GAN, specifically StyleGan2 by Karas [58], is used to generate them.

Tools and datasets

Table 2.2 shows an overview of the tools using different approaches for generating facial synthetic deepfakes using GANs and datasets shown in Table 2.2, the datasets section, which includes only synthetic images depicting the human face, which are used for training the discriminator.

Table 2.2: The table shows face synthesis tools with their publications and datasets. The table was taken from [41] and modified.

Tools	
Name	**Link**
clip2laten [83]	https://github.com/justinpinkney/clip2latentk
MMGeneration	https://github.com/open-mmlab/mmgeneration
StyleKD [113]	https://github.com/xuguodong03/stylekd
AdvFaces [31]	https://github.com/ronny3050/AdvFaces
ProGAN [57]	https://github.com/akanimax/pro_gan_pytorch
thispersondoesnotexist [58]	https://thispersondoesnotexist.com/
Datasets	
Name	**Content**
iFakeFaceDB	87,000 synthetic face images generated by the Style-GAN
TPDNE	60,000 face images
PGGAN	80,000

Usage and Attack vectors

This technology can be used for education, cinematography, health, and art; synthetically generated media would find applications in all the sectors mentioned. On the other hand,

the results of synthetically generated faces can be used for various disinformation campaigns, as well as for masking identities in cyber-attacks or fake profiles on social networks or for preserving anonymity in systems that require the upload of your photo. If state-of-the-art generators create realistic images, attackers can gain people's trust, leading to further abuse.

Detection

Faces generated by synthesis encounter inconsistencies and leave traces in certain areas. These critical areas are, for example, eyes, lips, and hair. An example is the color of the pupils because some GANs do not consider their color globally, and it is possible to notice color differences between the left and right eye. Also, the irregular shape and disparity of the pupils and the orientation of the reflections of the light source in the eye. Such differences would not occur in natural images. At present, the research focuses on more general approaches, such as orthographic detectors adapted to specific models, which show less performance for different sub-sets of data or modifications of the synthesized images. It is predicted that artifact-based methods will become ineffective due to the development of GAN models, which will remove these artifacts by their refinement. [116]

2.2.3 Face swap

This technique works by transferring one person's face to another person to make the resulting image as realistic as possible, as seen in Figure 2.8. For the first time, this concept was mentioned in 2004 by Blantz [12], and then a few years later, more advanced techniques were described by Bitouk in 2008 [11]. It can be a manual swap, but this approach is not considered a deepfake or a more common variant of swapping using GANs, NNs, and other techniques. The process involves aligning and mapping the facial features of two individuals and key intermediate steps, including tune of resolution, color, blur, and light [11]. Currently, the face-swapping approach is used mainly for entertainment purposes.

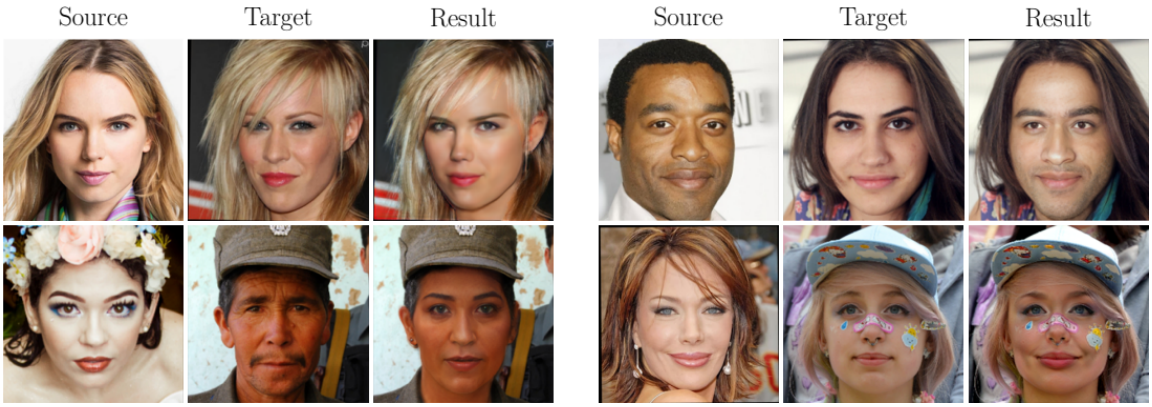


Figure 2.8: The figure shows examples of face swapping, where the facial features of the source person are digitally transferred onto the face of the target person, and the result is shown in the result face. The figure was taken from [118].

Some face swap approaches can suffer from a lack of memory when creating high-resolution images and instability in model training and data sampling, which can lead to worse results [76].

The swapping procedure is described in publications as follows [41].

1. **Extraction:** data is extracted from the source and target images, then processed by algorithms aimed at face detection, alignment, and segmentation.
2. **Training:** which results in better execution of face transposition from the source person to the target person.
3. **Conversion:** performing the actual transposition and retouching the result.

The most influential work in this area has used the Encoder-Decoder architecture for face-swapping. Another approach, in turn, proposes using a local-global approach, where the local branch involves adding local identity-related elements and trying to keep these elements as similar as possible in the target image. The local elements are facial details specific to each person: eyes, mouth, and nose. At the same time, the global branch adds elements of global identity, such as the shape of the head and cheekbones, and these elements are used to make the resulting whole fit. There are also currently many models that do not need to be trained before their use and access to such technology has been dramatically simplified for the average user, an example of which may be mobile applications that are easy to use even with limited device resources. [41]

Tools and datasets

The table 2.3 gives an overview of tools using different approaches to create deepfakes using face-swap. Also listed is DeepFaceLab[8], which implements a complete pipeline and is backed by a large community providing both support and pre-trained models. In Table 2.3 in the datasets section, examples of publicly available datasets containing images or videos of people swapping faces with faces are listed. Only datasets that, after subjective evaluation, show the higher quality of the resulting media have been selected. [41]

Usage and Attack vectors

This technology can be used for entertainment, as a face-swapping TikTok account, which uses the face of Tom Cruise [73], virtual reality, gaming, advertising, and historical recreation, as in the Salvador Dalí Museum in [5]; face-swapping technology finds valuable applications in all these sectors.

The main problem with this approach is the possibility of denigrating and abusing non-discerning persons for various political, criminal, or even pornographic and bullying purposes, as happened in New Jersey [106]. It is also essential to be cautious when forensic evidence may be falsified. Another possible misuse is an attack for identity theft and overcoming a system secured by poorly secured facial biometrics, which can be fooled even if this poorly secured system requires some user interaction, such as turning the head. [75]

Detection

With videos, detection is more straightforward because focusing on spontaneous and unwanted physiological activities such as breathing is possible. The creators of this content often overlook these activities because they are insignificant details, but they are essential in the detection process. In the case of images, detection is already more complicated since

[8] www.deepfakevfx.com/downloads/deepfacelab/

Table 2.3: The table shows face swap tools with their publications and datasets. The table was taken from [41] and modified.

Tools	
Name	**Link**
DeepFaceLab [81]	https://github.com/iperov/DeepFaceLab
GHOST [49]	https://github.com/ai-forever/ghost
One-Shot Face Swapping on Megapixels [123]	https://github.com/zyainfal/One-Shot-Face-Swapping-on-Megapixels
FaceSwapper [69]	https://faceswapper.ai/
FaceShifter [68]	https://github.com/maum-ai/faceshifter
MobileFaceSwap [115]	https://github.com/Seanseattle/MobileFaceSwap
Datasets	
Name	**Content**
Celeb-DF (v1,v2)	6000 deepfake videos
MFC Datasets	50,000 images and 500 deepfake videos
FaceForensics++	1000 original video sequences that have been manipulated with: Deepfakes, Face2Face, FaceSwap and NeuralTextures

there is little room for errors, and the difference between genuine and created images is often subtle and local. Despite this, artifacts are extracted from the image, simplifying the detection. This area also becomes interesting in the context of the extent to which people can detect a swapped face in an image. However, the findings show that with modern technologies, even the experts themselves have problems with detection without the use of detectors. [53, 41]

2.2.4 Face manipulation

Manipulation is widely used to modify facial features in a target image or video while preserving identity. Features such as hairstyle, eye color, skin texture, and overall age or gender can be manipulated, as seen in Figure 2.9. One of the most well-known applications are the filters on social networks, which allow for various manipulations ranging from adding glasses to the face to changing a person's gender and age. Manipulation is very similar to face reenactment, and no clear boundary separates them from each other. The approach focuses on more detailed adjustments that change various attributes of the face's appearance. The primary technology used by attribute manipulation applications is GAN. [29, 75]

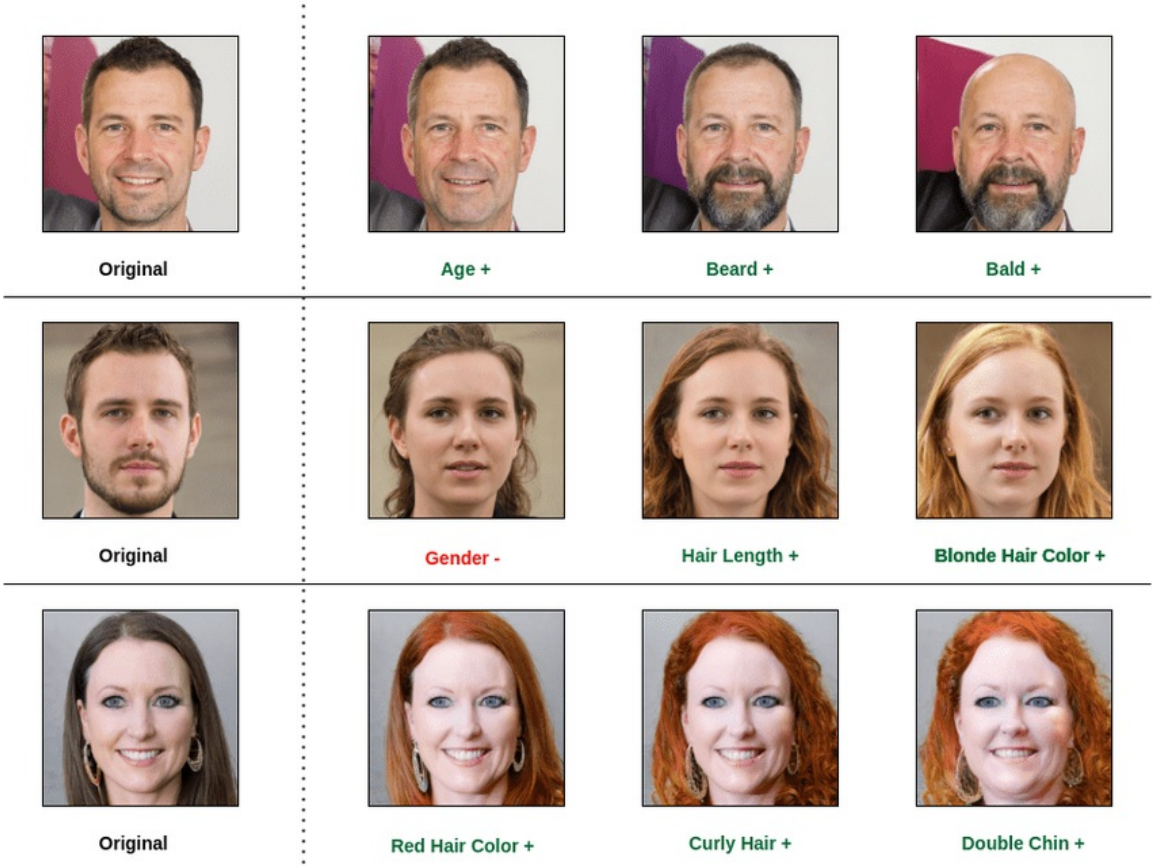


Figure 2.9: The figure shows three people whose facial features, age, gender, etc., were manipulated sequentially. The results show the final appearance of the person's face after manipulation but with minimal or no change in the other unmanipulated parts of the face. The figure was taken from [90] and modified.

Tools and datasets

The tools designed exclusively for handling insight are in a similar state to that of the reenactment. The available tools exist as open-source solutions with no additional user interfaces that require a certain level of user skills, but this does not apply to filters on social networks. The tools are listed in Table 2.4. In the case of datasets, there is a partial overlap with face reenactment datasets because most of the deepfakes with face reenactment include manipulation. As mentioned earlier, there is no clearly defined boundary between these approaches. The datasets are listed in Table 2.4 in the datasets section. [41]

Usage and Attack vectors

Applications that simplify attribute editing for ordinary users are widely used worldwide, but these applications need more results. More sophisticated methods are needed to create highly realistic edits due to the higher level of complexity. In terms of threats and attacks, manipulation can modify biometric facial features to deceive the biometric security system. Thus, an attacker can use manipulation to impersonate someone else, which can also have implications for crime and its associated evidence. If we were to compare the levels of threat

Table 2.4: The table shows face manipulation tools with their publications and datasets. The table was taken from [41] and modified.

Tools	
Name	**Link**
InterFaceGan [94]	https://github.com/genforce/interfacegan
StyleMapGAN [61]	https://github.com/naver-ai/StyleMapGAN
GAIA [91]	https://github.com/timsainb/GAIA
SURF-GAN [63]	https://github.com/jgkwak95/surf-gan
SkinDeep	https://github.com/vijishmadhavan/SkinDeep
Datasets	
Name	**Content**
Zhou et al. [122]	1005 deepfake images
Dang et al. [28]	240,000 deepfake images
FaceForensics++	1000 original video sequences that have been manipulated with: Deepfakes, Face2Face, FaceSwap and NeuralTextures

it poses to biometric systems, we would see that it is a more significant threat than face synthesis but less than face swapping and reenactment. [41, 29]

Detection

Such generated media is detected using approaches that focus on subtle irregularities in facial textures and physiological properties. Convolutional and recurrent neural networks, as well as GAN-embedded detectors, are used for detection. Among other things, human-assisted detection using heatmaps can also be used. [70]

Chapter 3

Deepfake Detection Techniques and Explainability

This chapter will review detection methods and explainability in deepfake detection. As mentioned earlier, deepfakes, in addition to their beneficial uses, are becoming a threat to the outside world in various areas. New technologies for creating deepfakes are constantly being developed; therefore, detection methods must keep up with them. Automated deepfake detection tools using machine learning are promising but face challenges due to the sophistication of generating fake content, which is often a step ahead. However, a limitation of this kind of detection is its significant non-transparency, which leads to poor interpretability and, hence, lower trustworthiness.

Explainability techniques such as heatmaps are being integrated into detection systems to address non-transparency. These techniques visualize on what basis the detection system arrived at a given result, dramatically increase confidence, and improve human understanding by highlighting the parts of the medium on which the system inferred the result. Explainability techniques also help improve user education regarding vulnerability to potential misinformation. Efforts to improve the explainability and control of AI-based systems are essential for the detection and the transparent use of other applications using advanced technologies. [25, 103, 114]

3.1 Overview of image Deepfake Detection Methods

This subsection provides an overview of general approaches to detecting deepfakes, focusing on manual and automated techniques. For both approaches, their strengths will be discussed, as well as the limitations and challenges they face.

3.1.1 Manual Detection Techniques

A person performs manual detection, and even if no guaranteed signs for detecting deepfakes have been defined so far, there are ways to detect manipulated media, but only to a certain quality. The content of this section is based on the work of Firc (2021) [40], Groh (2020) [48] and Johansen (2020) [55]. The detection includes the identification of subtle visual differences in the image, e.g.

Facial Features

- **Eyes**: This may be a poor rendering of pupils and irises. Pupils can be too large or, on the contrary, small or poorly shaped. Furthermore, detection is possible using light reflections in the eyes (Figure 3.1), which can be inconsistent or different in color (Figure 3.2).
- **Eyebrows and Shadows**: Eyebrows create and subsequently cast shadows that can be imperfectly projected onto the image (Figure 3.3).
- **Glasses and Glare**: The problem of accurately rendering the physical behavior of light and glare on glasses.
- **Facial Expressions**: Expressions unnatural to a person, such as the nose pointing in a different direction than the rest of the face.
- **Hair and Facial Hair**: Weakly rendered, possibly inconsistent and unnatural, often occurring.
- **Skin Features**: Irregularities in pigment shade, color mismatches between different body parts, unnatural or no birthmarks, and missing wrinkles.
- **Lips**: Inconsistency in symmetry, size, and color not matching the rest of the face.
- **Teeth**: Ambiguous contours, incorrect number, or even lousy shape of teeth (Figure 3.4).

Figure 3.1: An example of a face with missing light reflection in the eye in the target image. The figure was taken from [72] and modified.

Figure 3.2: A face generated using a GAN with a difference in eye color. The figure was taken from [72] and modified.

Figure 3.3: The figure shows incorrect shading caused by poor lighting estimation and the resulting nose geometry of a person in the resulting image. The figure was taken from [72].

Figure 3.4: The person in the resulting image has no tooth structure and is just a white spot without any contours. The figure was taken from [72].

General Indices

- **Visual Misalignment**: Blurry areas or misalignment, especially along the edges where the face and body meet (Figure 3.5).

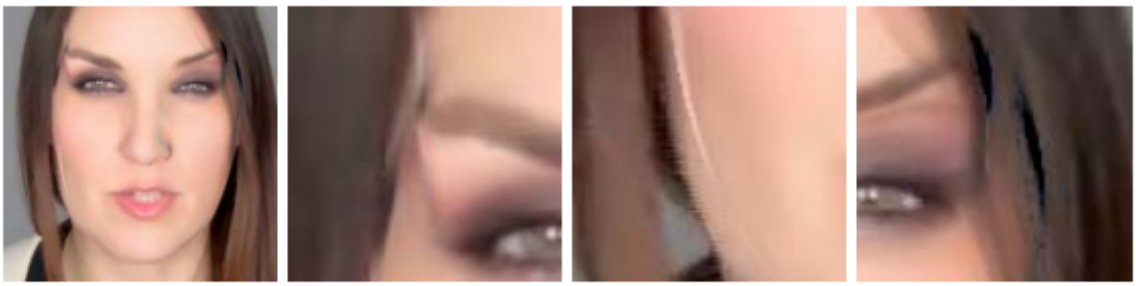


Figure 3.5: The image captures artifacts caused by incorrect geometry and alignment estimation. The figure was taken from [72].

These differences may raise the suspicion that the medium is a deepfake [40]. Humans are more accurate in face recognition compared to object recognition because they are naturally susceptible to facial proportions and deviations from the mean. However, human detection is not considered a scalable technique. It is also prone to fatigue and inattentiveness, which makes it increasingly difficult to detect deepfakes only with the eyes, and detection is thus considered challenging. Even experts can have a problem with correct

detection because subtle irregularities can easily be overlooked. On the other hand, even if quality fakes are hard to spot, it is possible to develop an intuitive sense of detection [42].

3.1.2 Automatic detection techniques

Automated deep image forgery detection techniques use computer vision, machine learning, attention mechanisms, and forensic analysis. Approaches mainly focus on identifying artifacts, lighting irregularities, and other facial image inconsistencies; these subtle imperfections can easily escape the human eye but are detectable using other computational methods. However, there are also techniques using frequency domain analysis to detect unique residuals that remain after image generation, mainly using GANs. This section describes an overview of some automated detection techniques and architectures and the ongoing challenges associated with different approaches to deepfake image detection. [4, 62]

Visual artifact detection

- **CNN-based models:** It is a fundamental building block of modern detection systems. They analyze the face region extracted from the image through multiple convolutional layers to detect patterns specific to fake images. This model focuses mainly on deformations and other anomalies found in the image, such as blurring, misalignment of some regions of the face, e.g., jaw, and inconsistencies at the pixel level, especially in the area of the eyes and mouth. An example of a widely used architecture of this type is EfficientNet [104], which demonstrates high detection accuracy on the FaceForensics++ dataset. Existing models are usually fine-tuned on additional labeled data to capture individual generative models' specific characteristics. [89, 97, 108]

- **Frequency-based analysis:** Methods of this kind try to detect by examining the frequency domain of the image, in which artifacts are often more evident than in the spatial domain. The generators introduce subtle irregularities into the high-frequency regions of the resulting medium. For frequency analysis, the images are therefore transformed using, for example, the Discrete Fourier Transform (DFT), which represents the image using the frequency components of which it is composed. Also, Gaussian Blur or Noise helps with the discriminating ability of fake images, such as in the spatial domain analysis of low-resolution images. High-pass filters, which can be used to detect fine-grained inconsistencies invisible to the eye, can also be an example of use. [119]

GAN Fingerprint Detection

- **GAN fingerprint identification:** GANs leave specific fingerprints, originating from the architectural limitations arising from the creation of deepfakes. Detection models, often using spectral analysis, can identify precisely these fingerprints and, based on them, decide whether they are fake. Sometimes, it is possible to identify precisely which GAN model was used to create, e.g., StyleGAN or ProGAN (Figure 3.6) because each has different specific characteristics. This approach is considered robust even in cases where traditional methods fail. [71, 80, 119]

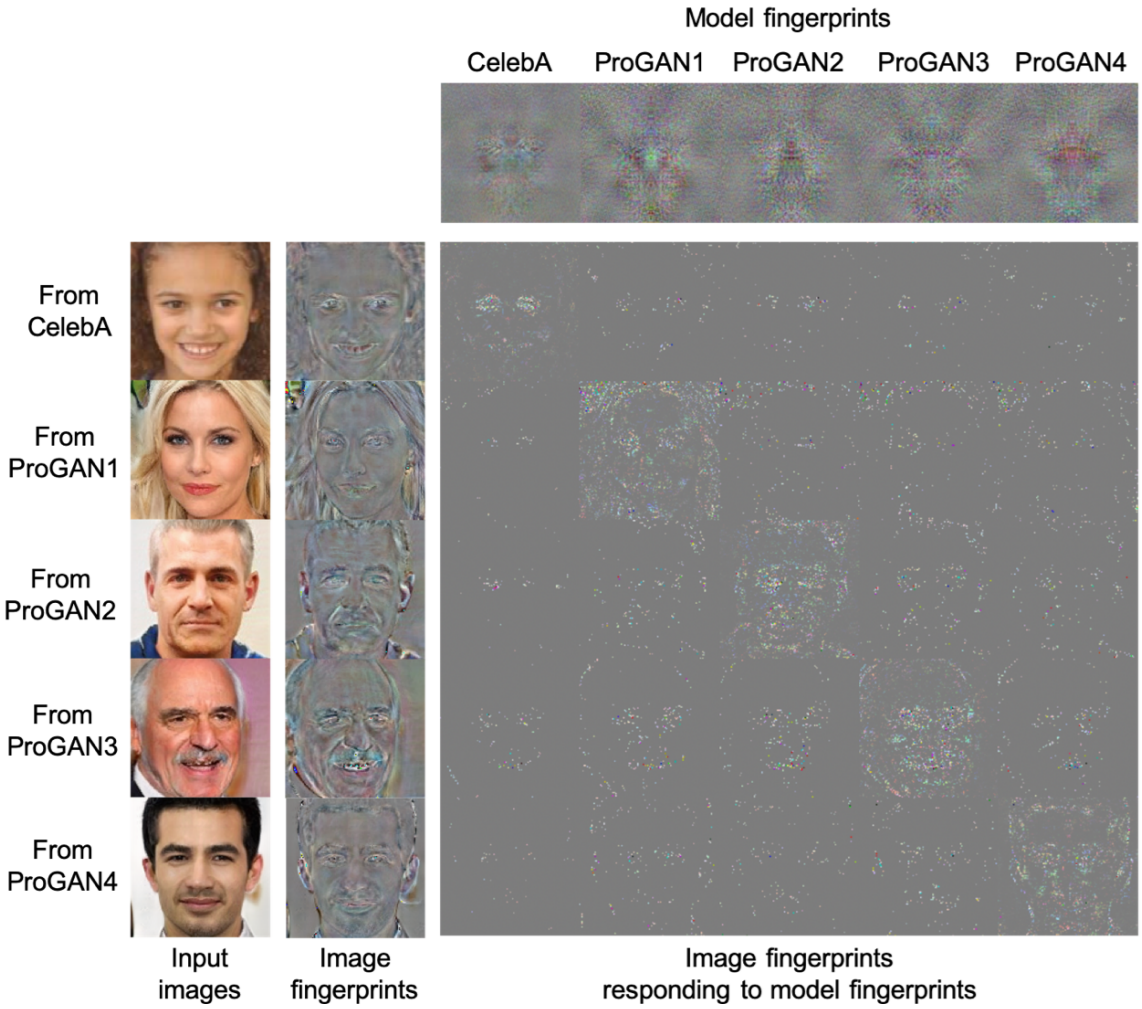


Figure 3.6: The figure illustrates the concept of model fingerprinting, which captures visual patterns from different generative models (ProGAN1, ProGAN2, etc.) and then compares them to input images. The resulting fingerprints display distinct characteristics of each and provide a way to identify the model used to create the image. The figure was taken from [119].

Contrastive Learning

- **Approach:** It is an approach that excels in exposing previously unseen methods, It focuses on training through commonalities between media of the same type, only genuine or only fake images, and also localizing differences between samples of different types, genuine together with fake images. It uses a score that is the basis of its operation and is a metric of the degree of similarity. The importance of this approach is also underlined by the difficulty of detecting deepfakes created by a different approach than the one on which the given model was trained. It tries to respond to the rapid progress in developing new technologies. The SimCLR framework [23] demonstrates the potential of contrastive learning. [23, 114]

Two-phase detection

The detection process is divided into 2 phases.

- **1. phase: General Features Extraction:** The general features of a person's face are extracted at a high level using pre-trained networks such as ResNet [14] or VGGFace [101]. Analyzing is based on the whole.
- **2. phase: Specialized models:** It uses smaller models focusing on inconsistencies and defects detected during the first phase. Analyzing based on smaller parts, therefore, reveals more specific imperfections. [79]

All the approaches mentioned are practical and try to identify deepfakes to the best extent. They aim to ensure generalizability so they work correctly even on previously unseen architectures, generating better results. However, they face several other challenges, such as the attacker's attempt to bypass automated systems and scalability, since in today's world whole of data, it is crucial to detect deepfakes in real-time and on a large scale, for example, in places such as social networks. To address these challenges, the priority should be to invent the most generalizable and robust solution possible. [6]

3.2 Importance of Explainability in Deepfake Detection

While deepfake detection methods can achieve high accuracy, it is equally critical to understand how the detector arrived at its result. The primary goal is to advance highly accurate deepfake detection, which also provides transparency mechanisms such as 2D heatmaps in the context of image deepfake detection [110].

There are several reasons why explainable AI is essential. Explaining its decision-making process is one of the critical capabilities of so-called explainable or interpretable AI. This is the first step towards achieving transparency compared to black-boxes.

Black-box models, created using deep learning algorithms, are very difficult to understand regarding how they work internally. This also applies to the authors of these black-boxes, who know the structure and weight settings of the model [2]. There is no known cut-off point when a model is considered a black-box.

Lack of interpretability is the leading cause of low confidence in the result. The use of explainable models solves this problematic and unknown behavior of models, often with several thousand to millions of parameters. However, it may involve a compromise in terms of lower performance in exchange for better interpretation.

By integrating explainability, the human user is provided with an overview and information about his decision-making processes; this feature thus leads to better understanding, acceptance, and trust in the outputs of the black-box model. The model's transparency helps its users understand why a specific decision was made, which is also helpful for better tuning and improving the system.

According to the European Union, e.g., GDPR, there are also regulations and laws that require the provision of explainability for automated decisions [111].

3.3 Heatmaps for Explainable Deepfake Detection

Research [120] has shown that not all pixels are processed equally during image processing by the detector because areas around the eyes, mouth, and facial contours can carry much greater informative value than other parts of the image.

Suppose any anomalies in the image indicate manipulation. In that case, the model can gradually, during training, learn to associate the anomalies with deepfakes and then use this knowledge to create heatmaps. This visual highlights areas in the image in 2D, using a color scale, with warmer colors indicating areas of higher importance [110].

This visualization technique allows for a better understanding of decision-making by highlighting critical areas of the image that had the most significant impact on decision-making during the deepfake detection process, which can be vital to improving the performance and credibility of the model [14].

Heatmaps can also guide a non-technical user when manually detecting deepfake and pointing out parts of the image to which they should pay attention.

In the case of a deepfake detector, the explainability of the heatmaps from 3.3 are based on the backpropagation of the gradient of the output class to the input image, and the heatmap is then finally represented as a weighted sum of all gradient values [96] or with various perturbations of the input image, more in the following subsections.

Techniques for Generating Heatmaps

There are several proven techniques for generating heatmaps, each with its methodology and strengths. This subsection describes some of the most commonly used approaches. Using these techniques helps improve the transparency and interpretation of black-box models, such as deepfake detectors, by highlighting the parts that most influence their decision-making process.

3.3.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Gradient-weighted Class Activation Mapping helps visualize models for any NN architecture without requiring architecture changes or retraining. The method applies to any type of NN architecture, making it highly versatile. [93]

It uses spatial information collected by individual convolutional layers to identify image regions that played a significant role during the classification process.

It involves calculating the gradients of the output of the last convolutional layer and highlighting the relevant parts through a map, which is a crucial step for the interpretability of the model. Gradients indicate the parts of the image that are most important for classification, and averaging them over all pixels yields the resulting weights. The calculated weights combine feature maps and thus the resulting heatmaps, where red indicates the most critical parts and blue indicates the least important. Heatmaps are displayed alongside the original image, making it easier for the observer to visualize the critical parts better. [84]

Grad-CAM is very suitable for comparing the performance of different deepfake detection techniques. The article [43] uses Grad-CAM, for example, to explain and reveal how deepfakes detection systems are fooled by the 2D-Malafid attack, and essentially shows how the attack can manipulate the focus and perception of the detection model.

However, Chattopadhyay [20] also introduced an advanced version, Grad-CAM++. Grad-CAM++ is an improved version that uses a weighted combination of first and second-stage output gradients, which leads to a finer final visualization and more accurate object detection and localization. [20, 21]

3.3.2 Local Interpretable Model-agnostic Explanations (LIME)

The principle of this technique is the local approximation of the model's behavior on a specific example by creating a more straightforward and interpretable model. For this, the input image is segmented, and its segments are then randomly masked, which makes the original image noisy. The noisy versions are then passed to the model, which tries to classify them. Finally, a simple linear model adapts the masks for each noise to the corresponding prediction scores. The resulting linear model weights then allow for visualization that reveals the extent to which individual segments influence the output of the detection model. In this way, it is possible to reveal which input had the most significant impact on the detector's decision-making. [110]

This approach is another superpixel visualization technique used to interpret predictions. For display, the original image is usually overlaid with a binary map or heatmap, representing the weight of individual parts [37]. It mainly points out essential areas of the face for detection, such as the eyes, mouth, and nose, and which the detection model focuses on the most. [84]

It is one of the most valuable and usable explainable tools for expressing any black-box complex model. It can also be helpful if the detection requires human verification of the result. [64]

3.3.3 SOBOL

It works similarly to LIME; the tool is model-agnostic and uses perturbation of multiple parts of the image. The goal is to quantify the importance of each feature for the decision score, individually and collectively [38]. Sobol-based sensitivity analysis is a global sensitivity analysis approach describing what uncertainty can be attributed to the output of a model based on various sources of uncertainty from the input. Global sensitivity analysis measures the sensitivity over the entire input space.

The basis is the use of a mathematical concept called SOBOL indices, which determine the degree of influence of the output variance based on individual randomly perturbed input variables. From a sequence of quasi-random numbers Quasi-Monte Carlo (QMC), sets of masks with real values are selected. QMC is used because of its more uniform and systematic distribution in space compared to random numbers. These masks are then applied to the input image via a perturbation function, e.g., noise and blurring; thus, perturbed inputs are created. The model then examines the inputs, which returns its prediction score. The obtained values and their respective masks are used for a visual explanation. These explanations are represented as heatmaps, created through Sobol-based sensitivity analysis, which describes the importance of individual areas using an estimate of the order of SOBOL indices. [45, 110]

The disadvantage of the tool is that it can only be used on image media, which is fine in the context of deepfake detection.

3.3.4 Randomized Input Sampling for Explanation (RISE)

Like SOBOL and LIME, RISE [82] works on randomly perturbing the input and observing the changes in the model's predictions. The perturbation in this technique is performed by generating binary masks, using Monte Carlo sampling, and applying them to the image, thereby creating a set of perturbed input images [38]. Analogously to SOBOL, these inputs are provided to the detection model, and a prediction value is returned for each input. The prediction values are used to weigh the importance of each mask. This method assumes that masks containing necessary image pixels have a greater weight than other masks [45]. The resulting heatmap is created by aggregating all weighted masks together. [110]

The disadvantage of this method is its higher computational time because high-quality visualization using a heatmap requires generating many masks. An insufficient number of masked input images results in temperature noise in random and unimportant parts of the image. Thus, the number of masks plays a crucial role in balancing the output quality and the computational time, which means that it is necessary to correctly set this tradeoff based on preferences. [50]

3.3.5 Shapley Additive Explanations (SHAP)

The SHAP method again uses superpixel features, which are gradually added to the input using Shapley features from game theory. The model attributes the effect of each input feature, sums up Shapley values fairly, and assigns rewards to each player based on their contribution to the outcome, which is then attributed to each pixel of the input image. Shapley values can be calculated efficiently when the dataset is small, but the computations become increasingly complex as the dataset grows. [92]

The article [110] compares the process to the coalition game, where the presence and absence of each participant impact the final outcome. The grand coalition results in an explainable map, where Shapley values distribute the contribution evenly across the coalition pixels. Model predictions for individual perturbation images are used to assess the importance of pixels.

SHAP can be used on any model because it is model agnostic. [44]

All techniques mentioned above can be seen in Figure 3.7.

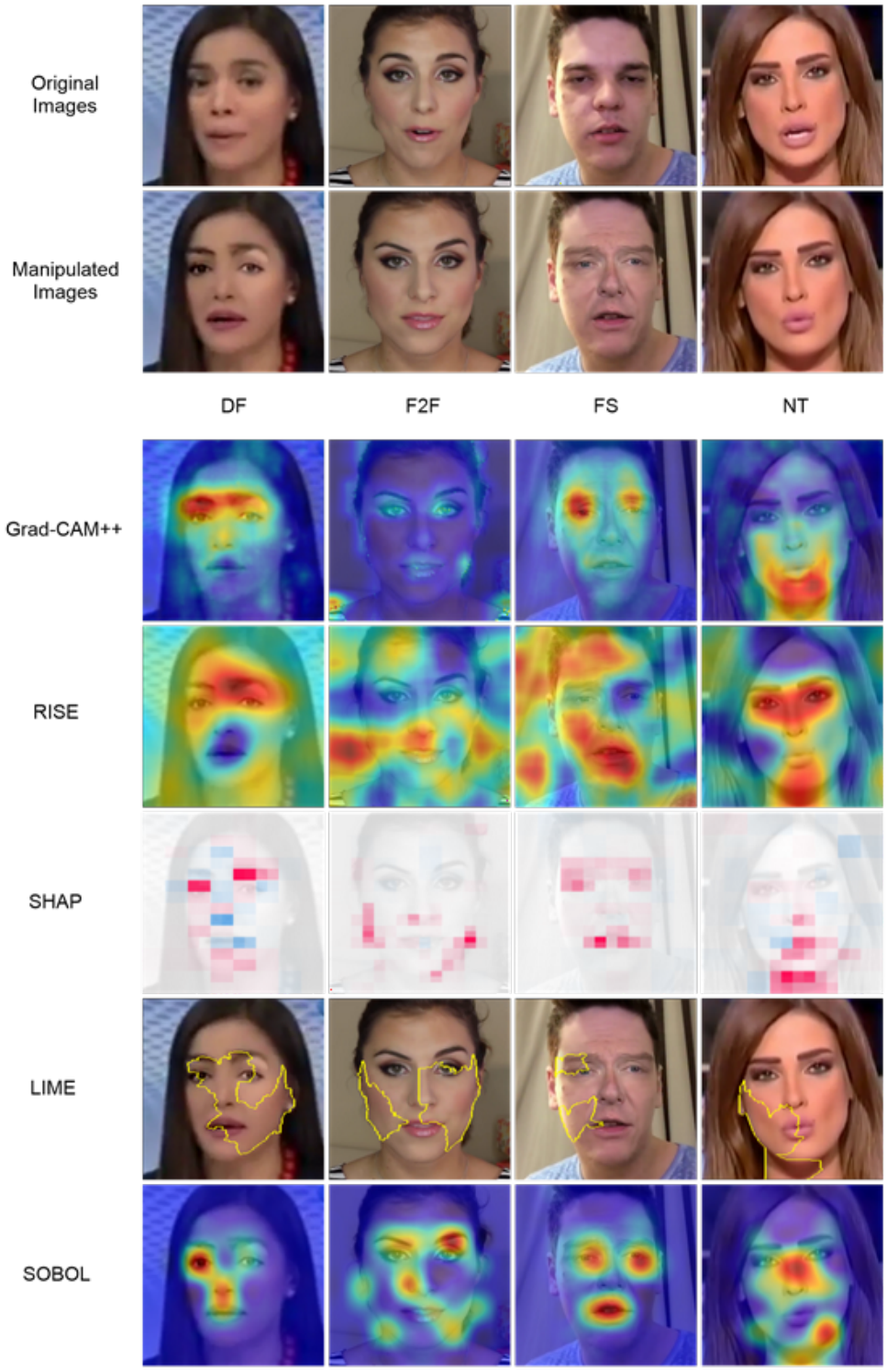


Figure 3.7: The figure shows how Grad-CAM++, RISE, SHAP, LIME, and SOBOL generate heatmaps to visualize the explainability of the deepfake detector for "FaceSwap" (FS), "DeepFakes" (DF), "Face2Face" (F2F), and "NeuralTextures" (NT). Pointing out the parts of the images on which the detector bases its decisions. The figure was taken from [110].

3.4 Challenges in Explainable Deepfake Detection

Regarding the explainability of deepfake detection systems, various challenges are mentioned that hinder their effectiveness and successful user adoption. The challenges are broadly divided into technical and user-centric challenges. Technical challenges focus on both the intrinsic complexity of the design and functionality. In contrast, user-centric challenges focus on the accessibility, credibility, and usability of user explanations. This section discusses the categories mentioned earlier in detail, highlighting specific obstacles for the challenges and describing features specific to robust and user-friendly systems.

3.4.1 Technical Challenges

Advanced deepfake generation techniques also bring various challenges associated with their detection and complicate the process of visualizing the decision-making of a deepfake detection system. Highly complex generative models, such as GANs and diffusion models, place significantly higher demands on the detection models themselves [102], and this has a significant share in their less effective transparency and interpretability. There are several technical challenges in the area of explainability [6]:

- **Model Interpretability:**

  Ensuring sufficient interpretability of detection models is critical to increasing trust and understandability. It is crucial to achieve a clear visualization of decision-making processes, thereby increasing user acceptance and trust in the machine learning system.

- **Feature Attribution:**

  Another equally important function is identifying features that contribute most to the model's prediction. This identification makes it easier to understand how the model works by understanding the factors that most influence the model's decision, thereby improving the identification of weaknesses and potential biases.

- **Context-Aware Explanations:**

  Providing contextual explanations ensures that individual model explanations are tailored to the specific circumstances of the decision-making process itself. Providing additional context to the user improves their understanding of the detector's result. This results in easier human cooperation and a better understanding of the AI model.

- **Scalability of Explanations:**

  This challenge mainly concerns the real-time explainability of detection systems. The processes of creating explanations, in this case, the generation of heatmaps, must be efficient and fast enough to be used in dynamic environments without reducing the demands on their accuracy and comprehensibility.

It is crucial to address the challenges mentioned above and strive to achieve them to the greatest extent possible.

3.4.2 User-Centered Challenges

User-centered challenges are equally important in explaining deepfake detection as technical challenges. Challenges of this type aim to ensure that explanations are accurate, understandable, usable, and accessible to different groups of users. There are several user-centered challenges in the area of explainability [6]:

- **User-Friendly Explanations:**

  The property of a user-friendly explanation means presenting the explanation to the user in the most intuitive way possible. Easy-to-understand decision-making processes are suitable for understanding even by a broader range of non-expert users. The result is users' more positive acceptance and adoption of AI models.

- **Trust and confidence:**

  Trust and confidence are essential to building user trust in the tool's effectiveness. Users can easily slip into a lack of trust in the results of the system if they cannot easily understand how the system arrived at a given result. The process of building trust requires the most straightforward and most consistent explanations possible. a model with such a feature results in the user gaining confidence in the decision-making process based on the clear and reliable explanations provided.

- **Cognitive overload:**

  One of the challenges of user-centered design is to provide enough information to explain the problem and avoid cognitive overload. It is counterproductive if a heatmap or other visualization tool contains less unstructured detail, which can overwhelm a non-expert. The goal is to find the right amount of information to provide enough information to easily understand how the model arrived at its result and draw conclusions from it. [110]

Chapter 4

Decision Support Systems for Deepfake Detection

The world we live in is driven by information, which often overwhelms us with a huge amount of data that exceeds the limits of information processing by the human brain. People are increasingly coming into contact with potentially misleading information that is misleading or false, and it is necessary to decide what to believe and what not to believe correctly. [67]

In this context, there is a need for tools that help individuals navigate large amounts of data or focus on specific parts of it during their decision processes. Tools will help them develop their critical thinking skills and enable them to distinguish fact from fiction. The solution to this problem may be advanced in artificial intelligence tools, through which users can more quickly and accurately evaluate the origin of information or media, e.g., image deepfake during decision processes.

The chapter describes the importance and potential of the systems and the challenges of use associated with assisted decision-making. It focuses on human-AI interaction, typically deepfake detectors, and describes methods to improve detection accuracy.

4.1 Definition and Role of Decision Support Systems (DSS)

The term DSS represents a concept in which a computer or other suitable technology can be involved in the decision-making process and contribute to decisions through the analysis of complex data [39]. It also provides structural support, allowing users to make more informed and better decisions in domains that often require easily overlooked information to make the right decision.

With the increasing challenges in DSS, traditional systems have gradually expanded, and various tools based on artificial intelligence are now increasingly integrated. While traditional DSS was based on predefined rules and static models, AI-DSS is dynamic and adaptive, which allows for solving more specific and difficult decisions. AI provides recommendations to a person, but the final decision is up to the user. AI-DSS is widely used, and people often do not even realize that they rely on decisions from such models, e.g., weather forecasting, advanced vehicle assistance systems, autonomous driving, voice assistants, and others [78].

Another example can be found in the medical sphere, where skin cancer is diagnosed similarly using AI; however, a specialist must determine the final verdict on the diagnosis.

This hybrid approach improved diagnostic accuracy compared to the approach where the doctor acted alone and did not use the DSS due to his low confidence in the AI's capabilities and ignored its helpful information in cases when he should not have done so [109]. The opposite problem may occur when the doctor stops thinking critically and starts to trust the AI overly [8]. When the AI reaches an erroneous conclusion, the expert relies on it without any thought or possibly prefers the AI's answer to his better conclusion, which he reached without its help. It is, therefore, essential to consider the AI's information and draw conclusions based on knowledge.

4.2 Challenges in Decision Support for Deepfake Detection

AI has the potential to significantly improve the decision-making process of deepfake detection by providing recommendations, but the functionality of this collaboration depends on specific challenges. Based on the examples described in Section 4.1, it is possible to point out the following challenges [98] associated with AI-DSS, which also apply to deepfake detection processes.

4.2.1 AI complements human abilities

The challenge requires that a person be able to distinguish situations in the decision-making process when he should and should not consider helpful information from the AI detector. It is expected that humans using AI to detect deepfakes will not achieve worse results than humans making decisions alone. Complementarity describes situations in which a human's decision-making performance in collaboration with an AI's decision-making performance exceeds their independent performance without using the other's knowledge [9]. Some studies confirm complementarity, while others claim that the human factor in the detection process is unnecessary. In various studies [112], it has been shown that AI offered advice that significantly exceeded the accuracy that a human evaluating alone would achieve. In one study [99], the AI-human pair even outperformed the AI-AI and human-human pairs. In cases where AI outperforms humans, it is always appropriate to follow its advice, but this raises the question of why humans should be further involved in decision-making.

In ideal scenarios, humans rely on AI in problematic spheres, which they know is more accurate, and on their judgment in spheres in which they know it is not so accurate.

This can also be achieved through explainability methods. The goal should be to reduce the burden on humans and improve human-AI complementarity, which is also facilitated by the independence between human predictions and predictions from a deepfake detector. On the other hand, even significant differences between AI predictions may not be perceived as valuable by humans [47].

4.2.2 Humans understand the limits of AI capabilities

An important factor is the alignment of the human mental model concerning the capabilities of a given detector. The aforementioned mental models shape the process of perception of AI, trust in it, and reliance on it by humans. In the context of deepfake detection, a person knows which races of the population or deepfake creation technologies the AI can correctly detect. Thus, a person can identify media poorly evaluated for AI in the decision-making process and introduce subjective risk assessment into this decision-making process. A key

element in this challenge is knowing when it is necessary to take control of the AI detector and, in these cases, take the information provided with a grain of salt.

Human dependence on AI varies in different contexts; in situations where life is at stake, it is more natural for humans to prefer humans over AI, while in situations such as deepfake detection, they may tend to prefer the decision of the AI detector.

Another problem related to this challenge is the so-called human aversion to the algorithm [18]. It occurs when a person is also familiar with the result of the validation of the detector's decision, which has shown that the detector was wrong in some cases [35]. In such a case, the person loses trust in the detector and returns to favoring human decisions, even if AI outperforms humans on average. However, aversion can be prevented by providing feedback and additional context, thanks to which the user is willing and able to use the information from the detector effectively.

4.2.3 Effective interaction between humans and AI

The challenge requires understanding the functionality of various designs aimed at mutual interaction between the detector and the assessor. The task is to develop accurate mental models to improve the efficiency of mutual communication. It focuses mainly on two main options: the timing of assistance during the decision-making process and the level of information provided so as not to overload the cognitive abilities of the assessor and his subsequent blind reliance on the detector.

According to the article [98], there are four options for the timing of providing auxiliary information, such as a heatmap and/or a result from a deepfake to a human:

- **Concurrent:** Auxiliary information is provided to the assessor from the beginning of his assessment of the image's origin.
- **Sequential:** a variant also known as the "Judge Advisor System" [13]. The auxiliary information is displayed to the assessor only after he has made his own decision and has the opportunity to update his previous decision afterward. According to [46], this variant also supports the independent reflection of the assessor.
- **On-Demand:** Allows the assessor to request help from the detector selectively. The approach is a variation of the sequential approach, both of which require the first step in the judgment to be made by a human before providing the auxiliary information and subsequently requesting/not requesting it.
- **Delayed and Temporary:** Delay provides delayed information to the assessor and thus provides him with the opportunity to use the initial time for his reflection on the image. The user is temporarily provided with additional information from the beginning, but only for a limited time. According to [85], there is a greater chance that a human will find the AI detector's errors if they have more time to examine the information provided.

The challenge is to choose the right approach to maximize the success of deepfake detection after the cooperation of the detector and the human assessor.

4.3 Explainability in Decision Support Systems

As mentioned in Section 3.2, explainability is an essential technology in assisted decision-making and is among the key enabling elements in compelling and credible deepfake detection.

Integrating explainable mechanisms into deepfake detection, in particular, through heatmaps described in Section 3.3, addresses various challenges associated with the transparency of black-box models.

Heatmaps provide visual feedback by pointing out the parts of the image that influenced the detector's result most and offer a clear overview of its reasoning process.

This increased transparency strengthens the AI-human collaboration process, described in Sections 4.1 and 4.2, leading to more informed decisions by leveraging the detector's strengths and the assessor's judgment.

Chapter 5

Experiment Design

The experiment aims to evaluate the impact of deepfake detectors and their explanation method "heatmaps" on the human decision-making process in classifying fake and genuine faces. Although state-of-the-art detectors currently achieve high success rates in image classification [36, 95, 110], when designing the experiment, it was consciously decided to take a slightly different approach than simply demonstrating the extent of the maximum potential for help in situations where deepfake detectors usually excel.

Stating that a scenario focused only on cases in which AI will provide highly reliable and correct answers in the vast majority of cases confirms the reliability and effectiveness of the detector itself, but it does not provide any meaningful information about the actual dynamics of human interaction with AI in less unambiguous cases. If the assumption is that there is a perfect detector with infallible certainty, the human factor would lose any justification at that point in the classification process. Therefore, the real value lies precisely in the aforementioned gray areas, i.e., in situations in which the deepfake detector is not always sure of its classification, the classification does not correspond to reality, or the detector made a mistake. These are cases that make up a very low percentage of current state-of-the-art solutions, but are crucial for a correct understanding of the level of trust, reliability, and human ability to critically evaluate and possibly correct AI outputs.

The experiment will focus on whether the outputs from the detector improve accuracy and certainty in classification and examine user behavior in the aforementioned gray area situations. The goal is to understand to what extent individual auxiliary information from the AI detector affects the judgment and trust of participants in the classification process, while the auxiliary information from the detector is not necessarily correct.

Research Questions

- **Research question 1 (RQ1):** To what extent can people correctly distinguish between genuine and fake images of a person's face?
- **Research question 2 (RQ2):** What is the impact on users' accuracy of providing a percentage score and a heatmap as outputs of a deepfake detector?

Hypotheses

- **Hypothesis 1 (H1):** Participants without support (*Control group*) will achieve lower accuracy than the deepfake detector group (*Detector group*) and deepfake detector + heatmap group (*Det. & Heatmap group*) when distinguishing between genuine and deepfake images.
- **Hypothesis 2 (H2):** The output of a deepfake detector will significantly influence the decision-making of participants in groups exposed to this information (*Detector group* and *Det. & Heatmap group*), leading them to align their decisions with the detector's output.
- **Hypothesis 3 (H3):** The explainable heatmaps visualizing the detector's decision-making will enable participants in *Det. & Heatmap group* to make more informed decisions, thereby mitigating blind trust in the detector's output compared to *Detector group*.

5.1 Methodology

This section will deal with a detailed description of all the elementary parts that make up the experiment, mainly the description of the experimental groups and participants, as well as further materials such as datasets or selected AI deepfake detectors and their training. This will be followed by a thorough description of the selection of candidate images for the experiment's needs and what information will be collected from the participants. Finally, the procedure and techniques that will be used to evaluate the experiment will be described.

5.1.1 Participants

The subsection describes the target group, which will constitute the vast majority of participants, and also describes the experimental groups and the minimum representation of participants in the given groups.

Target Group

The study focuses on a broad age representation with a likely predominance of young individuals over the age of 18.

Sample Size

A minimum of 45 participants will be recruited, divided into three experimental groups:

1. ***Control group*:** No assistance is provided. ($n \geq 15$)
2. ***Detector group*:** Provides the output percentage score of the deepfake detector, representing the probability of a deepfake. The score will be provided to the participant concurrently with the image; this approach was described in Section 4.2.3. ($n \geq 15$)
3. ***Det. & Heatmap Group*:** Provides the output percentage score of the detector and a heatmap visualizing its decision-making process. The score, which represents the probability of a deepfake, and the heatmap will be provided to the participant concurrently with the image, same as the previous group, described in Section 4.2.3. ($n \geq 15$)

5.1.2 Materials

This subsection will further specify the dataset used, its subsequent processing, and the training and use of selected deepfake detectors.

Dataset

A publicly available dataset FaceForensics++ (FF) [89] containing both genuine and deepfake videos of human faces will be employed (including only deepfake types Face2Face, FaceSwap, Deepfakes, NeuralTextures). Each dataset's video is pre-classified as genuine or deepfake.

Face images were extracted from videos with light compression by cropping frames. Specifically, 10 frames were selected for each video: the first and last and eight frames evenly spaced. Subsequently, faces within these frames were detected using MTCNN [121]. Importantly, the image was used only if the probability of detecting the face was higher than 95%; otherwise, another image was checked. Finally, each detected face was scaled to make the whole face visible, with 380×380 pixels as the final resolution, and the selected images were not further edited (e.g., resized, cropped, or color adjustments). This process used 52,000 images cropped from 5200 videos as a face images dataset for deepfake detector training.

For the training phase, the dataset was divided for each detector at the video level, in the same ratio of 80% training data (around 4200 videos, of which 42,000 images), 10% validation data (around 520 videos, of which 5200 images) and 10% test data (around 520 videos, of which 5200 images). It should be noted that the individual sets were different for each detector.

Deepfake Detectors

Two deepfake detectors were trained. The rationale for using two detectors is to increase the robustness of the results and minimize potential biases inherent in a single detector.

As the backbone architectures for the deepfake detectors, EfficientNet [104] models were chosen, specifically model B3 and model B4, which achieve perfect classification results for FF deepfake images according to [36, 95, 110], also seen at Figure 5.1.

Following this section, the detectors were pre-trained on ImageNet [32] and then trained on a training set with an epoch count of 100, a batch size of 32, and a learning rate of 0.001; these values were also used in [95]. The detectors were trained to classify two groups of genuine and fake images, where their output was a value representing the percentage probability of a fake image (e.g., a probability of 85% means that, according to the detector, there is an 85% chance that the input image is a deepfake). Subsequently, the B3 and B4 models were evaluated on their respective test sets. The evaluation results can be seen in Table 5.1, which demonstrates their high detection success rate.

According to Table 5.1, the trained deepfake detector models demonstrate their high success rate on the training set. However, the experiment was more concerned with cases that pose a challenge, even for the detectors themselves. Therefore, the selection of images used in the experiment was chosen to include cases evenly across the entire percentage spectrum for correct and incorrect classifications by the detector. The exact description and rationale for selecting images will be explained later.

Explainability methods via heatmaps were implemented to provide insight into the detector's decision-making process, with the Grad-CAM++ technology mentioned in 3.3

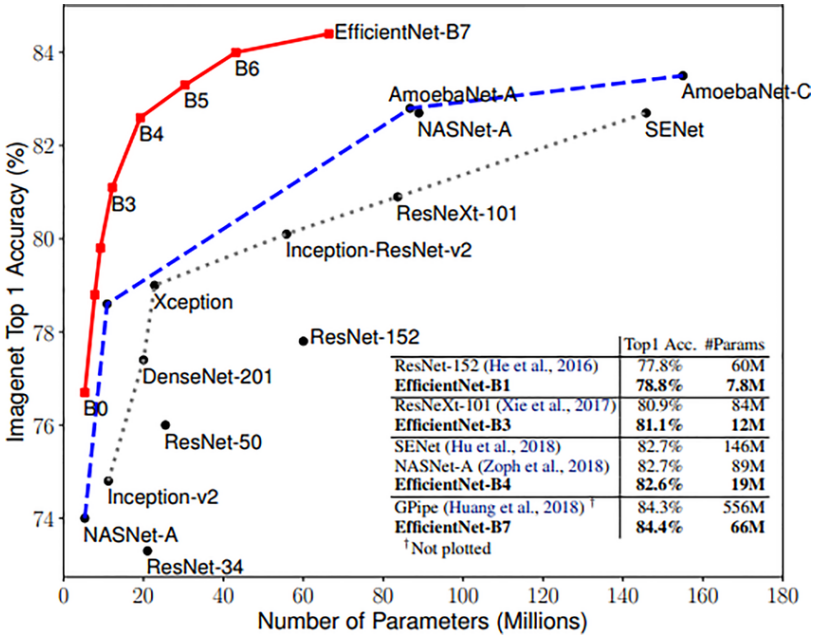


Figure 5.1: Model Size vs. ImageNet accuracy. EfficientNets significantly outperform other ConvNets. Figure was taken from [104].

Table 5.1: The table summarizes the performance metrics and the number of parameters in millions for the EfficientNet architecture models. Both models demonstrated high accuracy and Area Under the Receiver Operating Characteristic Curve (AUC) values, demonstrating their strong ability to discriminate between genuine and fake images.

Model	Params (M)	Test Loss	Test Acc (%)	Test AUC	Best Val AUC
EfficientNet-B3	10.7	0.4693	92.7	0.9634	0.9693
EfficientNet-B4	17.6	0.4495	92.4	0.9654	0.9639

chosen as the technology for both detectors, thus providing a second output from the detector. The great advantage of this approach is that it operates posthoc and thus was implemented directly on the selected detector without the need to change the internal structure or retrain the model.

The implementation of the explainability method was followed by a thorough empirical assessment of the detector results on the relevant test set, from which a set of candidate images for each detector was created. Further, based on the consensual decision of the author, supervisor, and consultant of the diploma thesis, 30 candidates were selected. This set will contain 15 genuine and 15 deepfake images of various types listed in the 2.2 section.

Indeed, this selection strategy is crucial as the experiment will focus on whether the outputs from the detector improve accuracy and certainty in classification and examine user behavior in the aforementioned gray area situations, which is why the confidence scores of candidate images are strategically distributed in the range of 0-100% to include low and high confidence values.

- **Score Distribution:** The distribution will be uniform, divided into five equal bins across the 0-100% range. This ensures an equal representation of confidence scores across the entire spectrum.

- **Gender Distribution:** Equal representation of the gender of people in the picture: 15 men and 15 women.
- **Correctness Mix:** Equal representation of correctly classified images and incorrectly classified images by the deepfake detector.
  - **Bin Specification for Each Detector (30 images total):**
    * 0-20%: 6 images (3 genuine, 3 deepfake, and of them 3 men, 3 women)
    * 21-40%: 6 images (3 genuine, 3 deepfake, and of them 3 men, 3 women)
    * 41-60%: 6 images (3 genuine, 3 deepfake, and of them 3 men, 3 women)
    * 61-80%: 6 images (3 genuine, 3 deepfake, and of them 3 men, 3 women)
    * 81-100%: 6 images (3 genuine, 3 deepfake, and of them 3 men, 3 women)

5.1.3 Procedure

This subsection describes the test platform on which the experiment was performed, further describes the user instructions before the start of the experiment, and provides a description of the individual experimental stages.

Testing Platform

The experiment took place in the form of an online questionnaire via the Limesurvey platform[1], and completing the questionnaire on a computer was recommended.

Pre-Experiment Instructions

Prior to the experiment, participants will receive detailed instructions, including:

- Information about their right to withdraw from the experiment at any time.
- a clear explanation of the experimental procedure.
- Information that all questions in the survey are optional.
- For *Detector group* and *Det. & Heatmap group*, guidance was provided on interpreting the outputs from the AI detector: a percentage score of the deepfake probability and a heatmap visualizing the detector's area of interest (only for *Det. & Heatmap group*). The instructions also warned about the detector's possible imperfection, but merely providing this additional information could bias their classification (the full text for *Det. & Heatmap group* guidance in Appendix A).

Experiment Stages

The experiment will consist of the following stages:

[1] https://github.com/LimeSurvey/LimeSurvey

Initial Questionnaire:

The introductory questionnaire consists of two types of question typologies.

- **Demographic Information:** These questions are focused mainly on the age group to which the research participant belongs, his/her gender, his/her highest level of education, and whether this education was technical. The exact wording of the questions and possible answers are attached in Appendix B.
- **Deepfake Experience:** Questions of this type are focused on the experiment participants' experiences with creating deepfakes, detecting deepfakes, subjective evaluation of their ability to detect deepfakes, awareness of the concept of XAI, their opinion on the impact of deepfakes on society, and also how often they come into contact with deepfakes. The exact wording of the questions and possible answers are attached in Appendix B.

Group Assignment:

Participants will be randomly assigned to one of the three experimental groups, ensuring a relatively equal distribution across groups.

Image Classification Task:

- Participants will be presented with the 60 pre-selected images in a randomized order.
- For each image, participants will be asked to make two judgments:
  1. **Binary Classification:** Is the image genuine or a deepfake? (Genuine/Deepfake)
  2. **Confidence Rating:** How confident are you in your classification? (Likert scale: Not at all, Slightly, Moderately, Very, Completely)
- Presentation based on group:

  The information provided to the participant depends on the group to which they are randomly assigned; this subsection shows an example of how each experimental group's selected sample image looks (Figure 5.2 for *Control group*, Figure 5.3 for *Detector group*, for *Det. & Heatmap group* Figure 5.4 and more Figures C.2, C.3, C.4 in Appendix C).

Figure 5.2: ***Control group*:** Only the deepfake image will be displayed.

Figure 5.3: ***Detector group*:** The deepfake detector's output score and the deepfake image will be displayed concurrently.

Figure 5.4: ***Det. & Heatmap group*:** The deepfake detector's output score, the deepfake image, and the corresponding heatmap will be displayed concurrently.

Post-Experiment Questionnaire:

These questions addressed participants' subjective feedback, focusing on their overall confidence in the accuracy of their classification, the level of difficulty in understanding the information provided, and their reliance on the AI assistance provided (score for *Detector group* and *Det. & Heatmap group*, heatmap only for *Det. & Heatmap group*), as well as qualitative insights into the impact of the detector and heatmap on their decision, concluding with an opportunity for general comments. The exact wording of the questions and possible answers are attached in Appendix B.

Early Termination:

If a participant chooses to end the classification task prematurely, they will be prompted to complete the post-experiment questionnaire before exiting.

Results Display:

At the end of the experiment, each participant will be shown their overall classification accuracy, which will be calculated as the percentage of correctly classified images out of the total number of images classified by the participant.

5.1.4 Data Analysis

This subsection describes, in particular, the process of cleaning the obtained data, quantitative metrics, and statistical tests that will be used to evaluate the experiment.

Data Cleaning

After data collection, it will be necessary to clean the data to ensure its quality and relevance to the research questions and hypotheses. The cleaning process will consist of:

- Removal of incomplete responses, all records from participants who did not complete all phases of the questionnaire, including the final questions, will be removed from the data. These partial responses would prevent a full-fledged data analysis.
- Removal of participants who classified 30 or fewer face images may lead to biased results.
- Removal of participants exhibiting patterns of rushed or too-long completion. Excessively short completion times, e.g., 10% fastest participants, as it is assumed that many participants will want to fill out the questionnaire sooner, not read the instructions, do not have to pay attention to the details that often decide on the correctness of the classification or just quickly clicked through the questionnaire and 5% slowest, where it is necessary to eliminate too long filling out the questionnaire caused, for example, by long interruptions during filling it out, this will help us ensure data quality.
- Open-ended responses will be reviewed and cleaned of irrelevant or nonsensical content.

Evaluation Metrics

In the case of quantitative performance evaluations, the standard and specific metrics will be used for each participant and individual experimental groups. In the context of metrics, deepfake images will be considered as a positive class and genuine images as a negative class. Metrics will be used:

- **Accuracy:** This is a basic metric expressing the total proportion of correctly classified images out of all classified images.
- **Precision:** Quantifies the reliability of a positive prediction. High precision means that if a participant classifies an image as a deepfake, there is a high probability that the image was a deepfake.
- **Specificity:** Measures the ability to correctly identify genuine images. High specificity means the participant can correctly classify genuine images and rarely classify genuine images as deepfake.

- **Recall:** Measures the ability to correctly identify deepfake images. High recall means that the participant can correctly classify deepfake images and rarely classifies deepfake as genuine.
- **F1-Score:** Represents the harmonic mean of precision and Sensitivity (recall). The output of the F1-Score is the only metric that balances both of these sub-metrics. Overall, it is a more robust assessment than accuracy.
- **Confidence-Weighted Score (CWS):** This is a metric designed specifically for this type of experiment, so that it is possible to take into account not only the accuracy of the classification, but also the measure of the confidence Rating obtained via the Likert scale appropriately. The CWS calculation will be as follows:
  - For each classification of a single image, a base score will be assigned based on the level of confidence (scale 1="Not at all sure" to 5="Completely sure"). confidence 1 ("Not at all") and any classification, whether deepfake or genuine, will be considered a zero-confidence guess and receive a base score of 0. confidence 2 to 5 will be converted to a base score of 0.25 to 1.0 (specifically: $(confidence - 1)/4$).
  - If the participant's classification of the image is correct (matched the image's context), this base score is added to the participant's total score.
  - If the classification is incorrect, this base score is subtracted from the participant's total score.
  - The final confidence-weighted score for the participant is the sum of these partial (positive or negative) scores for all 60 images. The minimum possible number of points is -60 and the maximum is 60.
  - Before calculation, participants who had a missing confidence response for 20% or more images were excluded from the analysis of this metric to help avoid biasing the results for participants who repeatedly ignored this part of the response during image classification.

  This metric rewards correct and confident decisions and penalizes incorrect and confident decisions, while uncertain answers neutralize the final score.

Additionally, the following analyses will be conducted:

- **Hypothesis Testing:** Appropriate statistical tests will verify the hypotheses and examine statistically significant differences. Due to the characteristics of the data, the assumption is that the primary use of non-parametric tests. Test for the evaluation:
  - Kruskal-Wallis test (for comparison of three independent groups).
  - Dunn's post-hoc test (for pairwise comparisons after a significant Kruskal-Wallis test).
  - Mann-Whitney U test (for comparison of two independent groups).
  - Spearman's correlation coefficient (for analysis of a monotonic relationship between two variables).
  - Friedman's test (comparing three or more dependent measurements).
  - Conover's post-hoc test (for pairwise comparisons after a significant Friedman's test).

  - Wilcoxon's paired test (for comparison of two dependent measurements).

  The choice of a specific test will depend on the specific analysis and the data.

- **Confidence Rating Analysis:** The distribution of confidence ratings will be analyzed to understand how confidence varies between groups and correlates with classification accuracy.
- **Time Analysis:** The time spent on the image classification page will be examined to see any patterns related to group assignment or classification accuracy.

5.2 Expected Duration

The estimated duration of the experiment is 10-25 minutes per participant. This will likely depend on the individual level of precision of each participant and their placement in one of the experimental groups, which will result in the possibility of a deeper analysis of the information provided.

5.3 Ethical Considerations

- All participants will be informed about the purpose of the study and their right to withdraw at any time.
- Data collected will be anonymized and used solely for diploma thesis and research purposes.
- Although it was not necessary due to the scope, conditions, and rules of the diploma thesis, the design of this experiment met the demands and requirements of the faculty's ethics committee, and its ethical aspects were officially approved by this committee, even in the case of further follow-up research.

Chapter 6

Evaluation of Experiment

This chapter will deal with the presentation and analysis of the experiment's results, which were described in detail in Chapter 5. The goal is to consistently analyze and quantify the data collected through an online experiment with three groups of participants using the Figures and metrics defined in the previous chapter.

Within the framework of the chapter, the first focus will be on the overall overview of the data, representing the number of individual experimental groups, presenting the demography, then overall accuracy metrics, and moving on to a detailed analysis focused on answering the stated research questions and also testing the stated hypotheses. Before the end, the focus will be on other interesting information that emerged from the evaluation process, and the conclusion will be a discussion that summarizes all key findings, their interpretation, the limits of the solutions and technologies used, and suggestions for further follow-up future research.

6.1 Participant Overview and Data Preparation

From March 2025 to April 2025, when data were collected on the LimeSurvey platform as part of the experiment, 260 questionnaire completions were collected, almost 6 times the required and expected number of respondents. However, it should still be noted that this is a rough sample before the data-cleaning process.

After applying the criteria described in the data analysis section 5.1.4, 13 research participants were first removed from the rough data sample due to non-completion of the questionnaire, another 14 were removed due to non-classification of more than half of the images in the central part of the questionnaire, and finally, 29 participant records were excluded from the data who belonged to either the group of fast finishers (under 350 seconds) or the group of slow finishers (over 2000 seconds). The final number of valid records included in the analysis was 204 participants after the individual cleaning phases.

To use the CWS metric, it is necessary in some cases to filter out participants who exceeded more than 20% of incomplete answers when classifying images to the question regarding confidence in their decision. The answers from 167 participants who met this criterion will be used in the evaluations using this metric.

The final numbers of participants after being assigned to individual experimental groups can be seen in Table 6.1. The representation in individual groups is relatively even, and there is no significant difference between the representations in groups for both the classical and CWS metrics.

Table 6.1: Participant Distribution per experimental group across two sets.

Group	Valid participants	Valid participants for CWS
Control	70	56
Detector	67	58
Det. & Heatmap	67	53
Total	204	167

6.1.1 Participant Demographics and Experiences

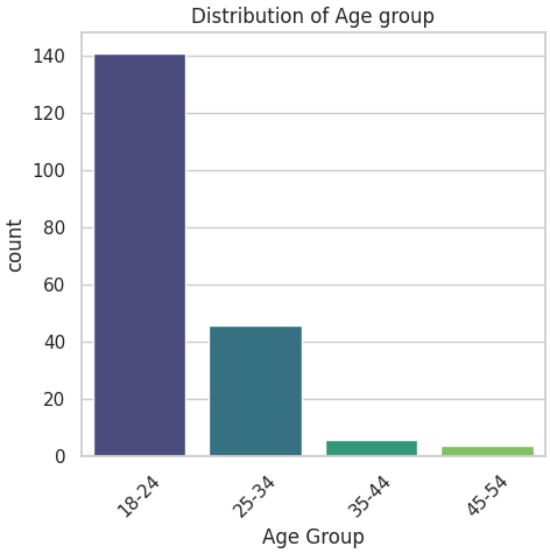
This section focuses on closer examination and understanding of the demographic characteristics of users and their experiences and knowledge in the field of deepfakes, which are relevant to this experiment, including a comparison of representation between experimental groups.

Demographics

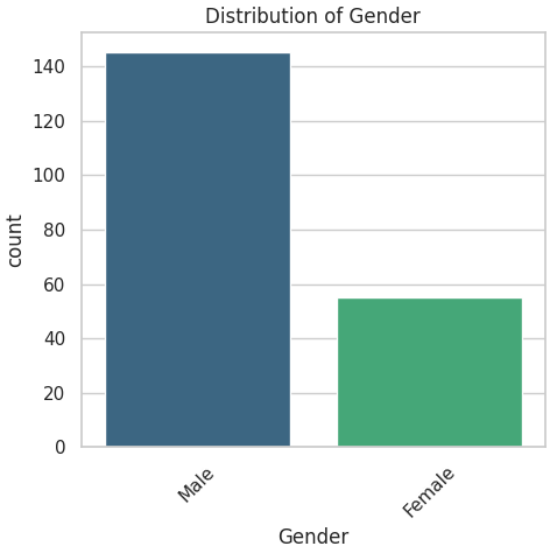
Within the demographic data, an analysis was performed according to age group, gender, highest education achieved so far, and whether it was a technical education.

- **Age group:** From Figure 6.1a, it can be seen that the vast majority of participants belong to the age group 18-24, and the next most represented is the group 25-34; these two groups group around 90% of respondents, which means that older age groups had significantly lower representation.
- **Gender:** As for gender, it is clear from Figure 6.1b that men predominated in the experiment, with about 70% of respondents representing them, compared to 30% representing women; this is since the survey was completed mainly by people from the faculty environment, which dominates the male population.
- **Education:** In the case of education, it can be concluded from the Figure 6.1c that the majority of respondents who completed the questionnaire were people with a current high school diploma, with a percentage representation of around 44%, followed by a group of participants with a Bachelor's degree, with 36%, a Master's degree was represented in 15% of cases. The group with a PhD degree had minimal representation. a question regarding technical orientation also addressed education. At the same time, Figure 6.1d shows that approximately 55% of people have a technical orientation, compared to 42% who do not, and the rest did not answer the question.

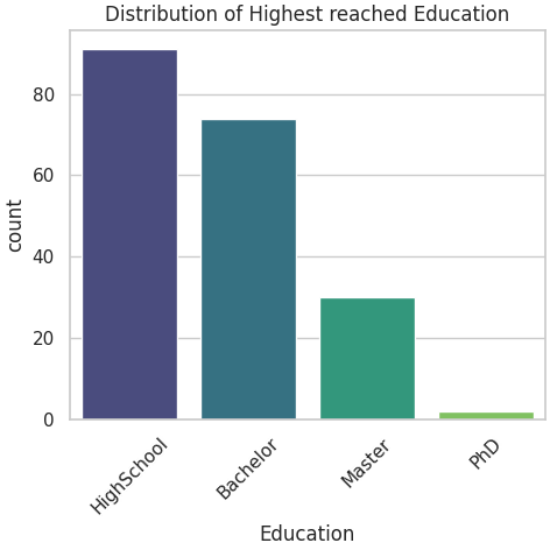
Random assignment of respondents to experimental groups ensured an even distribution of users between individual groups with minimal differences, which is also confirmed by Figure 6.2.



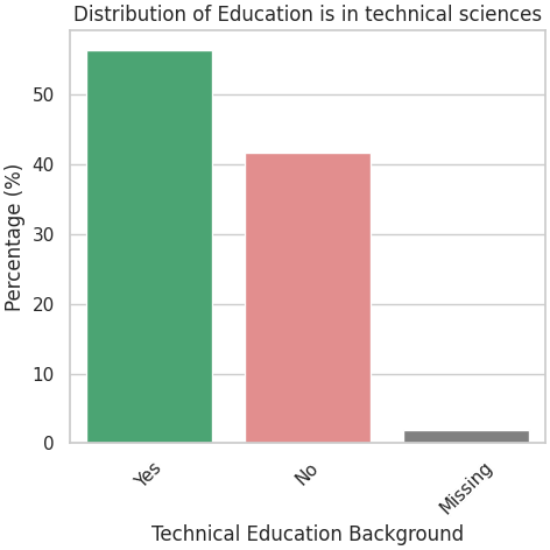
(a) Distribution of Age Group, which shows that the majority of representation is in the youngest group and then in the 25-34 age group, while other groups are minimally represented.



(b) Distribution of Gender, it is clear from the graph that male representation prevailed in the experiment.

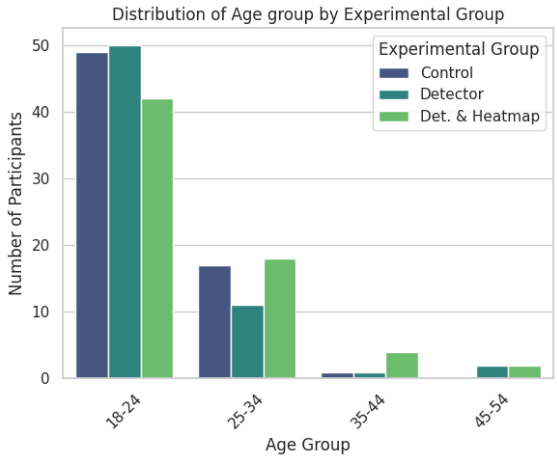


(c) Distribution of highest reached Education, the most represented are those with the highest education at the high school level, then Bachelor, Master and PhD have negligible representation.

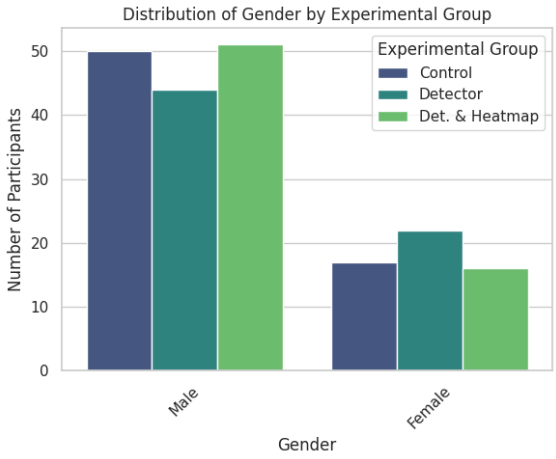


(d) Distribution of Technical Education Background (%), this is a more balanced representation, but users with a technically oriented education predominate.

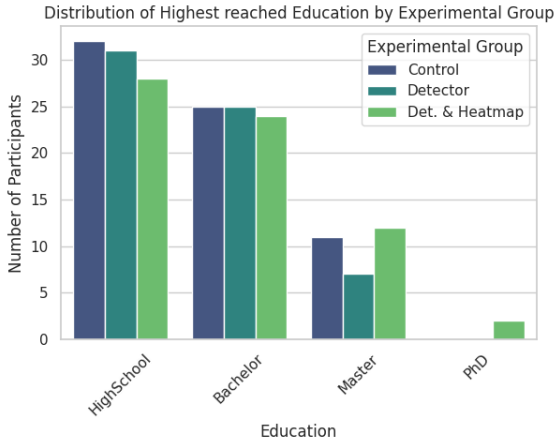
Figure 6.1: Participant Demographic Distributions, focusing on age groups, gender, and education.



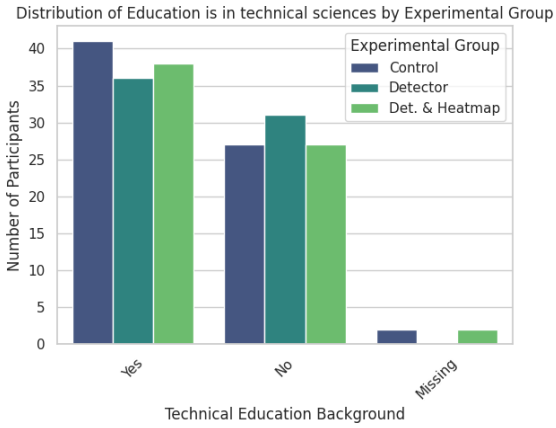
(a) Distribution of Age group between experimental groups.



(b) Distribution of Gender between experimental groups.



(c) Distribution of Highest Reached Education between experimental groups.



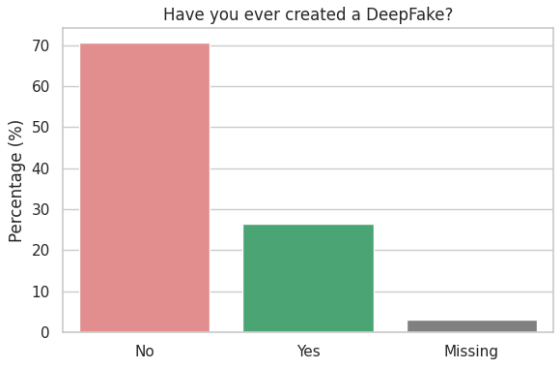
(d) Distribution of Technical Education Background (%) between experimental groups.

Figure 6.2: Participant Demographic Distributions between experimental groups, focusing on age groups, gender, and education. For each figure, there are apparently no differences between the demographic representation in the groups.

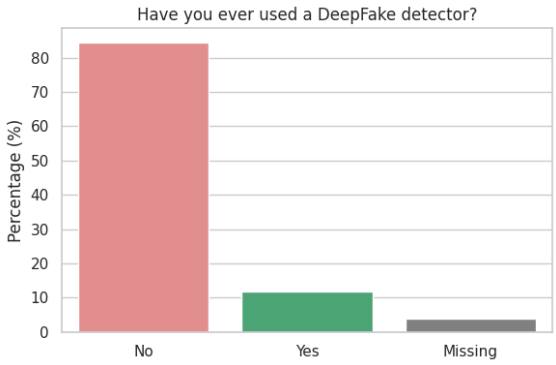
Previous Experience (Deepfakes, XAI)

In addition to demographic data, users were asked about their experience with deepfakes technology, detection, and awareness of XAI.

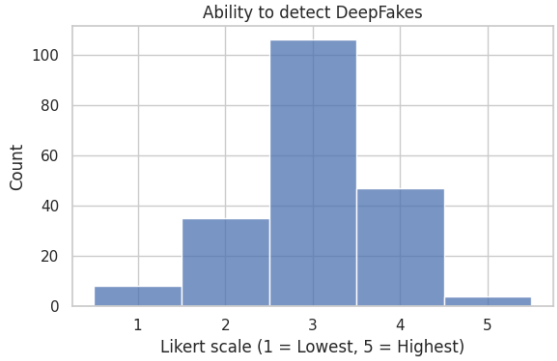
- **Deepfake creation:** Figures 6.3a and 6.3b show that approximately 70% of users have never created a deepfake, and more than 80% have never used any deepfake detection tool. These numbers indicate that the participants were more likely to be ordinary users than experts in this technology.
- **Deepfake detection:** When asked about the participant's subjective assessment of their ability to detect deepfake, Figure 6.3c shows that the majority of participants chose the value 3 ("Moderately") on a Likert scale from 1 ("Not at all") to 5 ("Completely"). Significantly fewer participants chose the other values on this scale. However, the Figure shows that people with the answer 4 ("Very") and 5 ("Completely") at least prevail over people with the answer 1 ("Not at all") and 2 ("Slightly"), which means that people are slightly optimistic about their subjective feeling about the ability to detect deepfake.
- **Explainable AI:** In the case of participants' familiarity with the term Explainable AI, it is clear from Figure 6.3d that most users have no idea or only a little knowledge of what it is actually about; the majority chose the value 1 ("Not at all") on the Likert scale and then the representation has a decreasing tendency up to the number 5 ("Completely"), which means that the participants have never encountered this term. Only about 17% of users chose the number 4 ("Very") or 5 ("Completely") on the scale. These people can be considered experienced in the field of AI.
- **Deepfake society impact:** When asked about the impact of deepfakes on society, it follows from Figure 6.3e that the participants consider the potential impact on society to be mostly high. On a 5-point Likert scale, over 50% of participants chose the value 4 ("Very"), which indicates significant impacts. Only a few people perceived the impact as Moderate or Slight, and no participant even chose 1 ("Not at all"), which means that everyone thinks that deepfakes have at least a minimal impact on society.
- **Frequency of deepfake seeing:** As for the frequency of how often people encounter deepfakes, it is clear from Figure 6.3f that every participant has come into contact with them at least once since no one chose the value 1 ("Never") on the scale. Furthermore, it is clear from the Figure that the number of responses in the middle and upper part of the scale prevailed, namely 3 ("Sometimes"), 4 ("Often"), or 5 ("Daily"), which indicates that users are aware of the presence of deepfakes and encounter them relatively often.



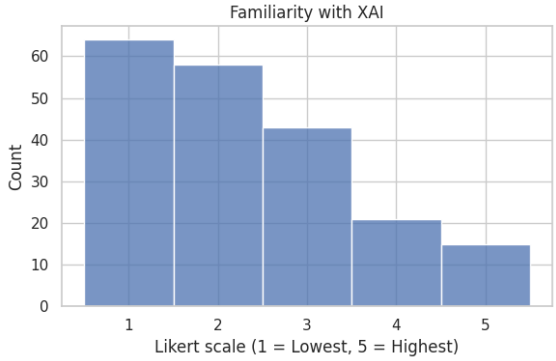
(a) From the figure, it is clear that a larger number of people participated in the experiment, around 70%, who had never tried to create a deepfake.



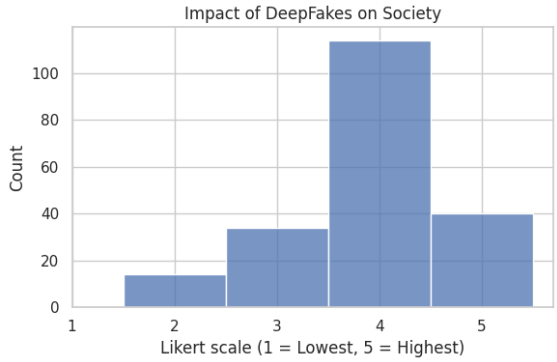
(b) More than 80% of the people in the experiment had never used any deepfakes detection tool. The number of more experienced participants in the field of deepfakes detection is only around 10%.



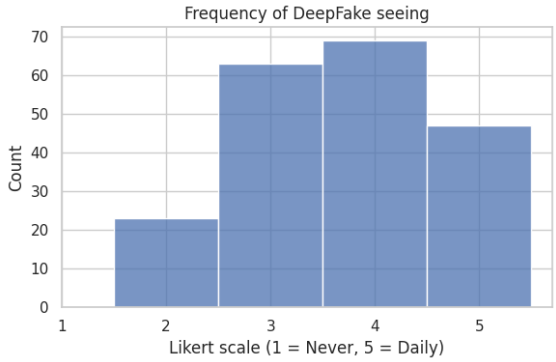
(c) The majority of the respondents chose 3 ("Moderately"), meaning a medium level of detection success. The other options indicate that people are slightly optimistic about their subjective feeling about the ability to detect deepfake.



(d) The vast majority of participants are not at all or only very slightly familiar with the term Explainable AI. Only 17% of people chose answer 4 ("Very") or 5 ("Completely"), which may indicate a representation of more experienced participants.



(e) Based on the Figure, the prevailing opinion is that deepfakes have a high impact on society. No participant thinks that deepfakes have no impact on society.



(f) It is clear from the Figure that only 11% of users do not encounter deepfakes often, however, the majority opinion is that they appear in society very often, even daily.

Figure 6.3: Distributions Related to DeepFake Experience and Perception.

6.2 Overall Performance Metrics

The section focuses mainly on the analysis of all important performance metrics: accuracy, F1-Score, precision, recall, specificity and Confidence-Weighted Score (CWS), and subjective self-confidence; all mentioned metrics are described in detail in Subsection 5.1.4. The goal will be to get a basic overview of how the participants of the experiment performed in the classification process and how they assessed the level of confidence in their decisions before proceeding to a detailed analysis of the research questions and hypotheses.

6.2.1 Classification performance

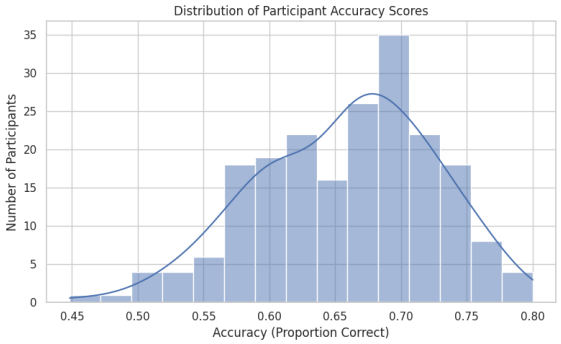
In this subsection, the basic metrics will be analyzed to assess the extent to which the participants can classify genuine and deepfake facial images.

Accuracy

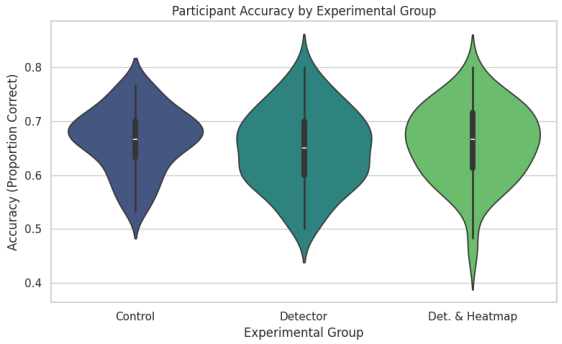
The metric represents the proportion of correctly classified images compared to all. In Figure 6.4a, it can be seen that most participants achieved an accuracy in the range of 57% to 75%, the average result achieved across groups was 66%, and the most frequent accuracy was 68%.

When focusing more closely on the classification accuracy between the individual groups, the violin graphs in Figure 6.4b show that their median accuracy reaches approximately the same values, and the overall distribution also appears similar. The mean and standard deviation values achieved by the individual group's *Control* (Mean = 0.666, SD=0.059), *Detector* (Mean = 0.648, SD = 0.072), and *Det. & Heatmap* (Mean = 0.659, SD=0.070) shows no sign of the difference between accuracy groups, which was also confirmed by the Kruskal-Wallis test, which reached p = 0.294, for alpha = 0.05.

The distribution visually appears to be normal, and based on accuracy, there is no statistical difference between experimental groups.



(a) Overall distribution of participant accuracy scores, showing a roughly normal shape centered around 0.65-0.70.



(b) Violin plots illustrating the distribution of participant accuracy scores for each experimental group, revealing similar medians and overall shapes across groups.

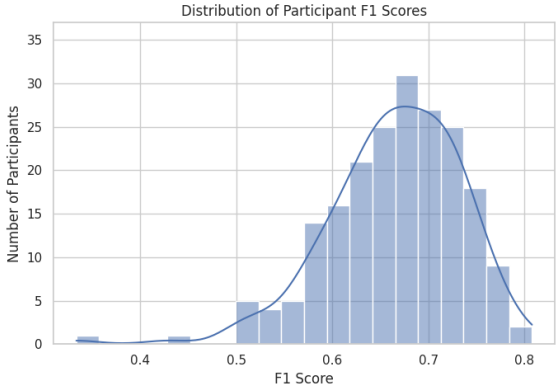
Figure 6.4: Analysis of Participant Accuracy Scores: (a) Overall distribution across all participants, and (b) distribution comparison by experimental group.

F1-Score

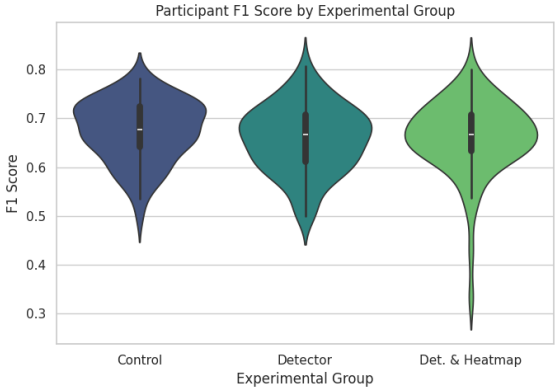
The metric is the harmonic mean of precision and recall and is considered a more robust metric than accuracy.

In Figure 6.5a, it can be seen that the distribution of achieved F1-Score across all participants appears to be normal, with two suspicious results at the level of 35% and 45%. However, overall, these are very similar to those shown by the accuracy graph, where the most common value was also around 68%, and the average was also at the level of 66%.

Upon closer examination of the achieved F1-Score between the individual groups, Figure 6.5b shows that they have almost identical median values. The mean and standard deviation values achieved by the individual group's *Control* (Mean = 0.675, SD=0.062), *Detector* (Mean = 0.658, SD=0.068), and *Det. & Heatmap* (Mean = 0.661, SD=0.076) show no indication of a difference between the F1-Score in the groups, which was also confirmed by the Kruskal-Wallis test, which reached p = 0.257, for alpha = 0.05.



(a) Overall distribution of participant F1-Score, exhibiting a similar roughly normal shape centered around 0.65-0.70 as seen for accuracy.



(b) Violin plots illustrating the distribution of participant F1-Score for each experimental group, indicating comparable medians and distribution shapes across the three groups.

Figure 6.5: Analysis of Participant F1-Score: (a) Overall distribution across all participants, and (b) distribution comparison by experimental group.

Other classification metrics

Other metrics were also analyzed to obtain a more comprehensive picture of the users' performance. Figure 6.6 shows the individual metrics and compares their performance between the groups. Deepfake images are considered a positive class, and genuine images are a negative class.

- **Recall:** Measures the ability to correctly identify deepfake images. As shown in Figure 6.6a, the median was 70% in the control and Det & Heatmap groups. The value in the *Detector group* was slightly lower, somewhere around 68%, but it indicates an almost equal ability to classify deepfake images correctly. The Kruskal-Wallis test reached p = 0.2242 for alpha = 0.05, meaning there is no statistically significant difference between the groups.
- **Specificity:** Measures the ability to correctly identify genuine images. The median ranged across all groups at around 62% according to Figure 6.6b and even in this

metric, no statistically significant differences between groups were found using the Kruskal-Wallis test, p = 0.6711, for alpha = 0.05, which is the value based on which can be said that the metric is far from a significant statistical difference across groups.

- **Precision:** Quantifies the reliability of an optimistic prediction. The medians ranged around 65% according to Figure 6.6c, again without significant differences between groups with p = 0.297, for alpha = 0.05 using the Kruskal-Wallis test.

At the overall level, the participants' performance between the experimental groups using the metrics above did not differ significantly.

6.2.2 Subjective Confidence and CWS

Another factor that can play a role in performance is the participants' confidence in their decisions, so these responses were incorporated into the performance metric and examined separately.

Average Confidence

Users rated their confidence in their decisions for each image on a Likert scale from 1 to 5 (Not at all, Slightly, Moderately, Very, Completely).

Figure 6.7a shows that participants' average confidence values were most often in the range of 3.0 and 4.0, with a peak around 3.5. The distribution appears symmetrical. A comparison of the average confidence between groups is shown in Figure 6.7b. However, again, there are no apparent differences at first glance, as confirmed by the Kruskal-Wallis test, where p = 0.875 for alpha = 0.05.

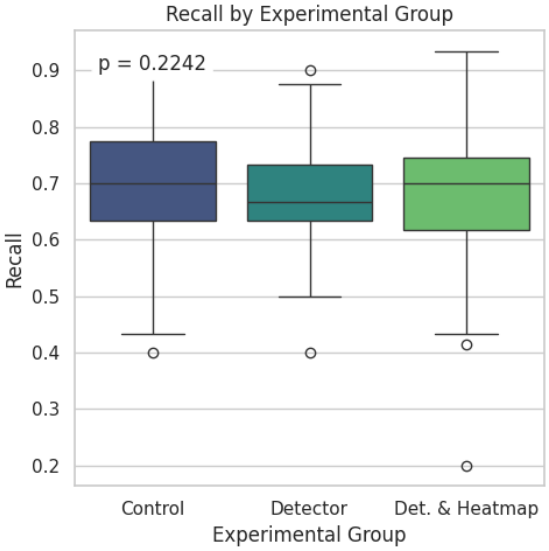
Confidence-Weighted Score (CWS)

a metric, which is precisely defined in Subsection 5.1.4, composes the classification of responses together with confidence, rewarding correct and confident responses and penalizing incorrect and overconfident responses. A smaller sample was used to analyze this metric since not all respondents met the specified conditions; the final number of respondents included in this metric is 167.

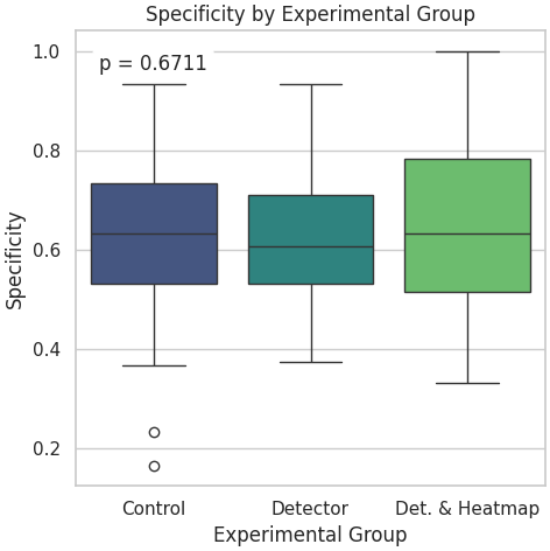
From Figure 6.8a, it is clear that the distribution lies in the interval -5 to 30, with a peak around 15. The predominance of mostly positive values means that users made correct primary decisions with more confidence.

Upon further examination of Figure 6.8b, it was concluded that there is no significant statistical difference between the groups for this metric, as confirmed by the Kruskal-Wallis test where p = 0.231 for alpha = 0.05.

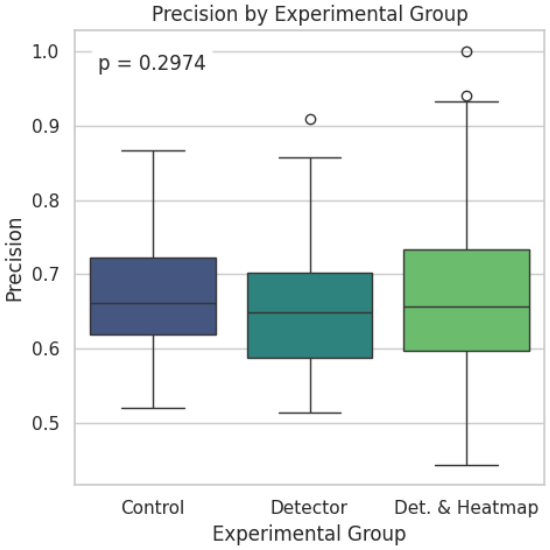
Although the graphs may indicate minor differences at the group level for these metrics, they were not large enough to be statistically significant. The following sections will focus on a more detailed analysis of these metrics in the context of the stated research questions and hypotheses, including pairwise comparisons and analysis of specific conditions.



(a) Distribution of participant recall scores by experimental group, measuring the ability to identify deepfake images (Kruskal-Wallis p=0.2242) correctly.

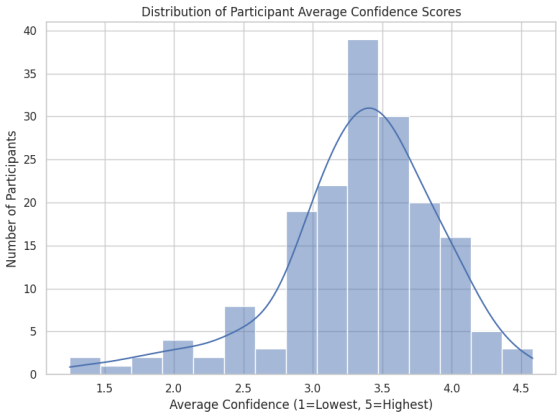


(b) Distribution of participant specificity scores by experimental group, measuring the ability to identify genuine images (Kruskal-Wallis p=0.6711) correctly.

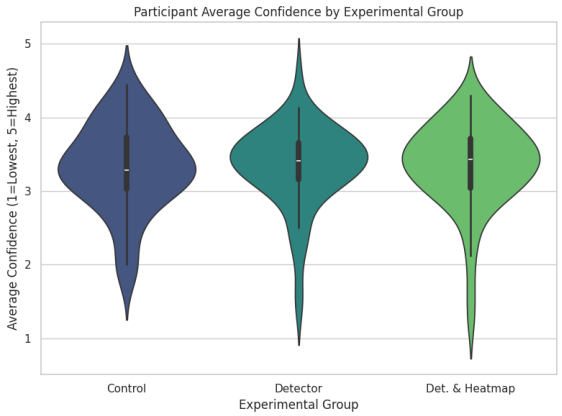


(c) Distribution of participant precision scores by experimental group, measuring the reliability of 'deepfake' classifications (Kruskal-Wallis p=0.2974).

Figure 6.6: Comparison of recall, specificity, and precision metrics across the three experimental groups using box plots, with Kruskal-Wallis p-values with aplha 0.05 indicating no significant differences between groups for any metric.

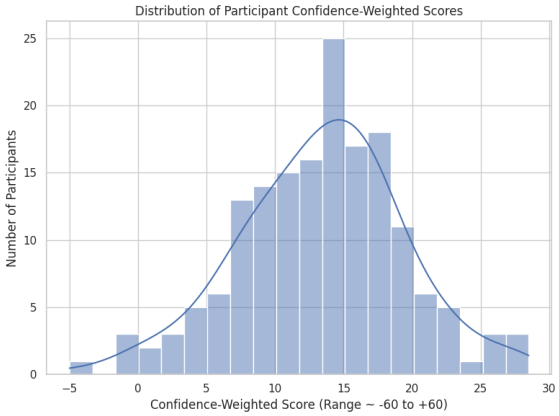


(a) Overall distribution of participants' average confidence ratings across all trials, centered around 3.5 ('Moderate' to 'Very' confidence).

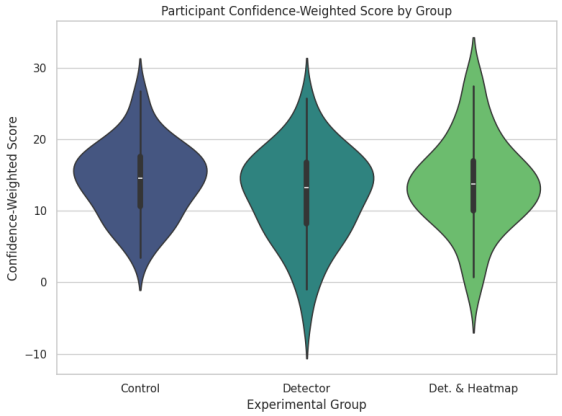


(b) Violin plots comparing the distribution of average confidence ratings across experimental groups, showing similar shapes and medians.

Figure 6.7: Analysis of Participant average confidence ratings (1-5 Likert scale): (a) Overall distribution, and (b) distribution comparison by experimental group.



(a) Overall distribution of participant CWS, ranging from approximately -5 to +30 with a peak around +15, indicating generally positive performance when confidence is factored in.



(b) Violin plots comparing the distribution of CWS across experimental groups, revealing similar medians and shapes.

Figure 6.8: Analysis of Participant Confidence-Weighted Scores (CWS): (a) Overall distribution, and (b) distribution comparison by experimental group.

6.3 Analysis of Research Questions and Hypotheses

The following chapter analyzes the collected data in more detail to test the established hypotheses and answer the research questions. It will gradually analyze the basic human ability to detect deepfake (RQ1), the impact of AI assistance in the form of percentage score and heatmap on performance (RQ2, H1), to what extent participants followed the suggestion from the AI (RQ2, H2) and finally, it will be analyzed what role the heatmap plays in decision-making and possibly mitigating blind trust in the percentage output from the AI detector (RQ2, H3).

6.3.1 RQ1: Baseline Human accuracy (*Control group*)

RQ1: To what extent can people correctly distinguish between genuine and fake images of a person's face?

To answer this question, the responses of people in the *Control group* (70 participants), i.e., people who classified the images without help, were analyzed.

Many metrics will be used to answer this question have already been analyzed in the Section 6.2. The performance of this group was not bad, but not by any means highly accurate. Based on the analysis in the section above, it was concluded that the mean accuracy among the *Control group* reached a value of 0.666 with a standard deviation of 0.059 and a median of 0.667 (Figure 6.4b). For F1, the median value in the group was also reached at the level of 0.667 (Figure 6.5b). Both median values are significantly above the level of 0.5, corresponding to random guessing, and therefore, it can be confirmed that the participants have a particular ability to distinguish deepfake images.

The relationship between the recall and specificity metrics was evaluated as part of a more detailed analysis. The result of a pairwise comparison of the selected metrics for each participant using the Wilcoxon signed-rank test is a significant statistical difference ($W = 752.5$, $p = 0.024$, $alpha = 0.05$), the median recall within the group reached a value of 0.7 while the median specificity was at the level of 0.633, which implies that participants who were assigned to the *Control group* were statistically more successful in correctly classifying deepfakes compared to correctly classifying genuine images.

Further the relationship between the recall and precision metrics was analyzed. Again, pairwise comparisons of the selected metrics for each participant were used using the Wilcoxon signed-rank test, with the result ($W = 771$, $p = 0.049$, $alpha = 0.05$), which confirms the statistical significance between the median recall at the level of 0.7 and the median precision at the level of 0.662. Higher median levels in the case of recall tell us that participants from this group had a higher tendency to classify images as deepfakes, which led to a higher rate of False Positives, which always increases when genuine images are labeled deepfakes.

Conclusion for RQ1:

Humans, without any help from AI, can correctly distinguish between deepfake and genuine images at the level of 66% (accuracy) and 67% (F1-Score), which is better than random classification as they can correctly classify 2 out of 3 provided images however their performance is not balanced as they are more successful in identifying deepfakes (recall) than genuine (specificity) and there is a statistically significant trend of more frequent classification of images as deepfake even at the cost of a higher number of errors in classifying

genuine images. This finding may also be a consequence of the fact that users were warned about what kind of experiment it was, and perhaps they were more biased towards detecting fakes during classification.

6.3.2 RQ2 & H1: Impact of AI Assistance on Performance

RQ2: What is the impact on users' accuracy of providing a percentage score and a heatmap as outputs of a deepfake detector?

H1: Participants without support (*Control group*) will achieve lower accuracy than the deepfake detector group (*Detector group*) and deepfake detector + heatmap group (*Det. & Heatmap group*) when distinguishing between genuine and deepfake images.

The first step to answering this research question is to test Hypothesis 1, which is based on the assumption of favorable outcomes caused by the outputs of the AI detector, whether the percentage output or the addition of a heatmap, compared to the group without AI assistance.

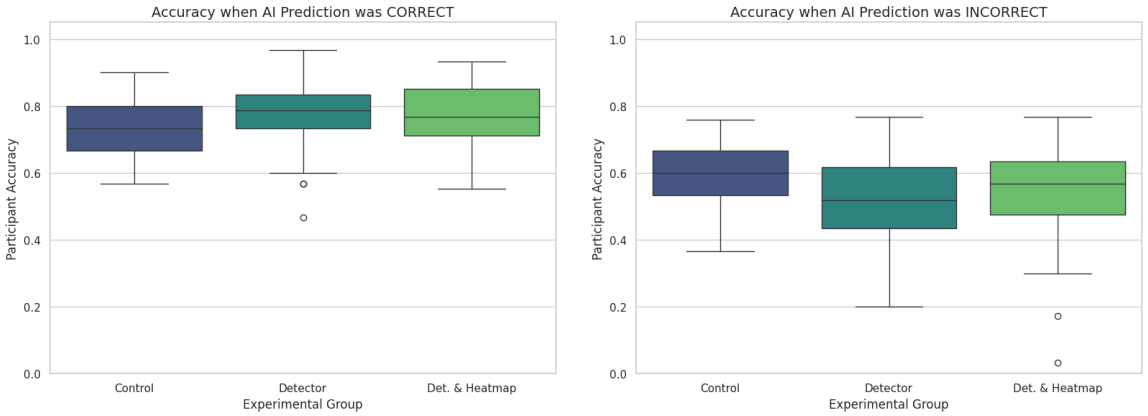
The comparison across groups for the accuracy and F1 metrics have already been analyzed in Section 6.2, where the Kruskal-Wallis statistical test yielded results with for accuracy $p = 0.294$ and for F1-Score $p = 0.257$, and at the significance level $alpha = 0.05$, there was no statistically significant difference between the experimental groups (Figures 6.4b and 6.5b).

In order to test Hypothesis 1 more thoroughly, a pairwise comparison of performance using the F1 metric were performed between the *Control group* and both remaining groups using the non-parametric Mann-Whitney U test:

- ***Control* vs *Detector* comparison:** The selected test achieved a result of $U = 2713$, $p = 0.114$, $alpha = 0.05$ and did not show a significant difference in F1-Score between the two tested groups.
- ***Control* vs *Det. & Heatmap* comparison:** The selected test achieved a result of $U = 2584$, $p = 0.304$, $alpha = 0.05$ and did not show a significant difference in F1-Score between the two tested groups.

However, it is important to note in the context of H1 that a specifically set experiment with demanding classifications and gray zones can mask participants' performance levels in individual groups. Therefore, an attempt was made to focus on the set of images for which the AI correctly predicted; this subset of images can simulate the case of a highly accurate detector. From Figure 6.9a, it is clear that the distribution and median value in *Control group* is optically significantly lower, somewhere at 72% compared to *Detector group* and *Det. & Heatmap group*; *Detector group* and *Det. & Heatmap group* have medians somewhere at 78%. This visual difference was also confirmed by the Kruskal-Wallis test, which revealed a significant difference between the groups ($H = 8.55$, $p = 0.014$, $alpha = 0.05$) and subsequent Dunn's post-hoc tests specified significance in performance (*Control group* vs *Detector group*: $p = 0.024$; *Control group* vs *Det. & Heatmap group*: $p = 0.036$; $alpha = 0.05$), *Detector group* and *Det. & Heatmap group* achieving statistically better accuracy than *Control group* in the case of a robust detector. The non-significance of the overall performance may be due to the performance of the AI in cases where it is classified

incorrectly (Figure 6.9b), which will be further investigated in the context of H3.



(a) Participant accuracy when the AI prediction was correct, showing higher median accuracy in the assisted groups (*Detector*, *Det. & Heatmap*) compared to the *Control group*.

(b) Participant accuracy when the AI prediction was incorrect, showing lower median accuracy in the *Detector group* compared to the *Control group* and *Det. & Heatmap group*.

Figure 6.9: Participant accuracy comparison across experimental groups, conditioned on the correctness of the AI's prediction for the presented image.

Conclusion for H1: None of the Paired tests showed statistical significance ($p < 0.05$). Therefore, the collected data do not meet the assumption that AI assistance led to a statistically significant improvement in F1 in difficult cases and gray area cases, and therefore, **hypothesis H1 is rejected**. However, AI assistance proved significantly beneficial in accuracy in cases where the AI prediction was correct.

Relation to RQ2: In the context of the experimental design, which focuses mainly on complex cases and also on gray area images, it can be stated that in these cases, even though participants *Detector group* and *Det. & Heatmap group* had additional information from AI, any statistically significant improvement in *Detector group* and *Det. & Heatmap group* compared to *Control group* wasn't observed. However, AI assistance proved significantly beneficial in accuracy in cases where the AI prediction was correct. A possible reason why it is like that may be the equally represented ambiguity and inaccuracy of AI outputs in a specific image set. This may lead to reduced effectiveness, misinterpretation, or participant distrust of these outputs. Therefore, further testing the other stated hypotheses, H2, and H3, is necessary to investigate how participants worked with this additional information and whether there is any trend in the different use of additional information between *Detector group* and *Det. & Heatmap group*. The investigation of the impact of the heatmap will be the subject of analysis for H3.

6.3.3 RQ2 & H2: Alignment with detector output

H2: The output of the deepfake detector will have a significant impact on the decision-making of participants in the groups exposed to this information (*Detector group* and *Det. & Heatmap group*), leading them to align their decisions with the detector output.

In this hypothesis, the assumption is that participants in *Detector group* or *Det. & Heatmap group* will not ignore the additional information provided by the AI and will follow it to some extent.

In a first attempt to test this hypothesis, was decided to quantify it using a match rate, which describes the proportion of trials in which the participant's response **matched** the AI prediction (prediction was defined as: AI score less than 50% = "Genuine" prediction, AI score greater than or equal to 50% = "Deepfake" prediction), with the AI prediction being incorrect in half of the cases. A high level of agreement would indicate a strong tendency of the participant to lean towards the classification from the detector.

When examining the measure, it was found that the *Detector group* had a Mean of 0.625 and a standard deviation of 0.089, comparable to the values for *Det. & Heatmap group*, where the Mean was 0.611 and the standard deviation was 0.098. From the box plots in Figure 6.10, it can be seen that the medians were around 63% in both groups, while a decreasing trend in agreement with AI can be observed in *Det. & Heatmap group*, which has a slightly lower median value. However, no statistical significance difference was confirmed by the Mann-Whitney U test with the result U = 2488.0, p = 0.279, alpha = 0.05.

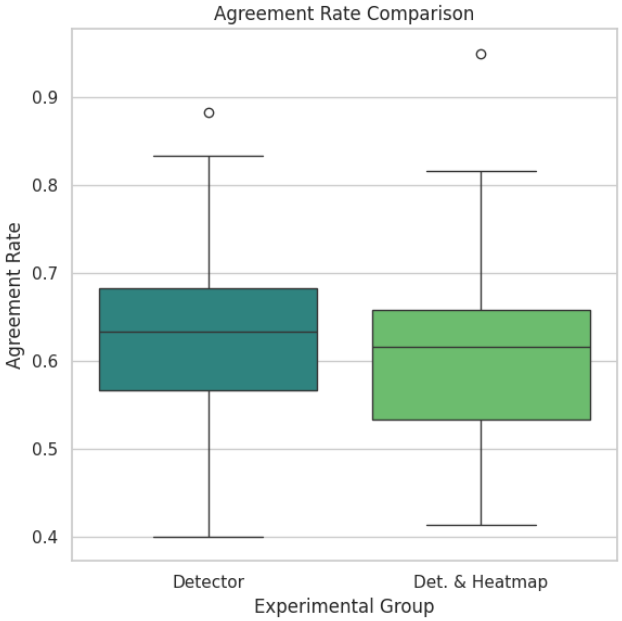
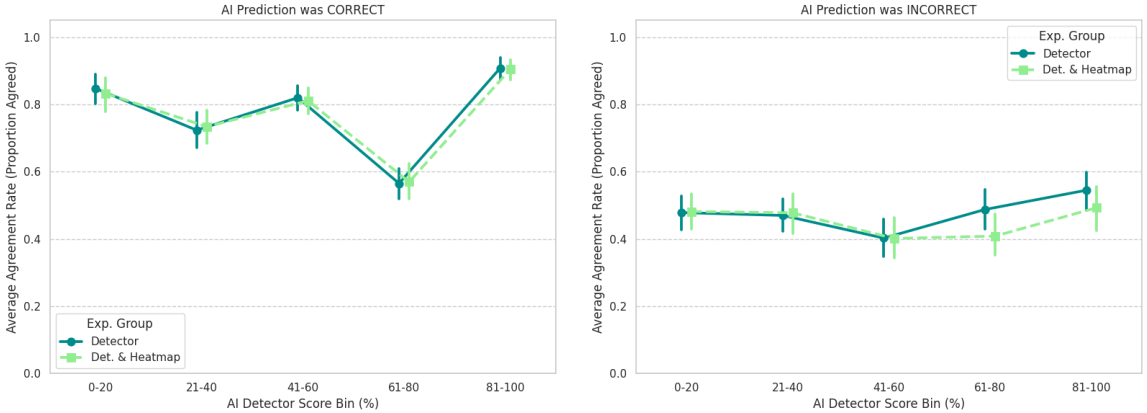


Figure 6.10: Box plot comparing the overall Agreement Rate (proportion of trials where participant decision matched AI prediction) between the *Detector group* and *Det. & Heatmap group*, showing similar median rates.

A more detailed and fascinating look at Figure 6.11 describes the change in the user agreement rate depending on the AI's confidence expressed as a percentage score and its prediction accuracy.



(a) Agreement rate when the AI prediction was correct, showing high agreement (esp. in high-/low confidence bins) for both groups, with similar trends.

(b) Agreement rate when the AI prediction was incorrect, showing substantially lower agreement overall compared to correct predictions, with both groups exhibiting similar rates across most bins.

Figure 6.11: Alignment Rate vs. Binned AI Score divided into five equally broad and deep bins, left for images where the AI predicted correctly (a), right for images where it predicted incorrectly (b). A percentage lower than 50 is considered genuine, and a percentage of 50 and above is considered a deepfake prediction.

Figure 6.11a shows that in cases where the AI achieved high confidence and correct prediction (0-20 – AI is sure that it is genuine, 81-100 – AI is sure that it is a deepfake) the user agreement rate with the AI reached 85% to 90%, however, a high agreement rate was also recorded in bins 21-40 and 41-60, around 80%, the only significant drop occurred in the 61-80 group, where the agreement rate dropped to less than 60%. Between *Detector group* and *Det. & Heatmap group*, no apparent trend in the graph would indicate differences between groups.

In cases of incorrect AI prediction (Figure 6.11b), there is a significant drop in the agreement rate with the AI, with all groups falling within the range of 40% to 55% regardless of the percentage prediction from the detector, with bin 81-100, representing a very confident and incorrect prediction, being able to influence some percentage of participants towards misclassification. An important finding, however, is that in cases of incorrect AI prediction, users more often disagreed with this prediction compared to predictions when the AI was successful, which means that they were often able to identify and correct the AI error, which suggests that participants in incorrect cases are not always blinded by their trust in the AI result. In the case of bins 61-80 and 81-100, a slight deviation trend can be seen; this will be investigated when testing H3.

Conclusion for H2: Based on the overall agreement rate data and a more detailed overview of the agreement rate divided into correct and incorrect AI predictions from Figure 6.11, the AI detector significantly impacted users' decision-making process in *Detector group* and *Det. & Heatmap group*. The agreement rate for correct AI predictions was high,

especially when the detector's confidence was high. On the other hand, in cases where the detector was wrong, participants were not completely passive and were often able to correct this incorrect classification. Hypothesis H2 is **supported** based on the data.

Relation to RQ2: Although hypothesis H1 was rejected due to the lack of evidence of a statistically significant difference between groups based on accuracy (accuracy and F1-Score), confirmation of H2 is an important part of RQ2, which deals with the overall behavioral impact of AI assistance since, based on H2 it is clear that the AI detector has a measurable impact on the decision-making process of the participant who worked with the provided score and, especially in cases where the AI signaled high confidence, was influenced by this score. The investigation of the impact of the heatmap will be the subject of analysis for H3.

6.3.4 RQ2 & H3: Role of Heatmaps in Mitigating Blind Trust

H3: The explainable heatmaps visualizing the detector's decisionmaking will enable participants in *Det. & Heatmap group* to make more informed decisions, thereby mitigating blind trust in the detector's output compared to *Detector group*.

Based on the hypothesis, the assumption is that heatmaps will help users make more informed decisions. In the context of the experimental design, the decision was to investigate cases where the AI predicted incorrectly and compare *Detector group* to *Det. & Heatmap group*, and whether, in the case of *Det. & Heatmap group*, the decisions were more informed and could correct the AI. These two groups will also be compared with the results achieved by the *Control group*.

In Figure 6.9b, there is a noticeable difference between the median level of *Control group* (60%) and the median of *Detector group* (52%) and *Det. & Heatmap group* (57%). Kruskal-Wallis test (H = 12.768, p = 0.0017, alpha = 0.05) and subsequent Dunn's post-hoc test confirmed the suspicion and revealed statistical significance between *Control group* and *Detector group* at the p = 0.001 and alpha = 0.05 levels. Statistical significance was not confirmed between *Detector group* and *Det. & Heatmap group* (p=0.170, alpha = 0.05). These findings indicate that providing only the score for incorrect predictions significantly impacts the deterioration of performance compared to the *Control group*. However, this deterioration can be mitigated using the heatmap and thus maintain the accuracy achieved by the *Control group*.

Next, in testing the hypothesis, focus was moved on the CWS metric; the mean CWS was calculated for each incorrectly predicted image (N = 30) and each experimental group. The median from this set describes the typical performance of the group. The Friedman test for differences between groups, which is suitable for comparing multiple dependent groups (CWS for the same images in different groups), was used. The test reached statistical significance (Chi-squared = 14.067, p = 0.0009, alpha = 0.05). The subsequent Conover post-hoc test with Holm's correction for multiple comparisons revealed that the *Detector group* achieved, in poorly predicted cases, significantly worse results than the *Control group* (p=0.0003, alpha =0.05) and also worse results than the group that also had the heatmap (p=0.048, alpha =0.05). The median in the individual groups was 0.108 for *Control group*, -0.031 for *Detector group*, and 0.048 for *Det. & Heatmap group*. The results again confirm that heatmaps help participants correct erroneous AI results.

In Figure 6.12, the trend can be found in the heatmap group, where this group has a higher rate of correct override of the detector prediction.

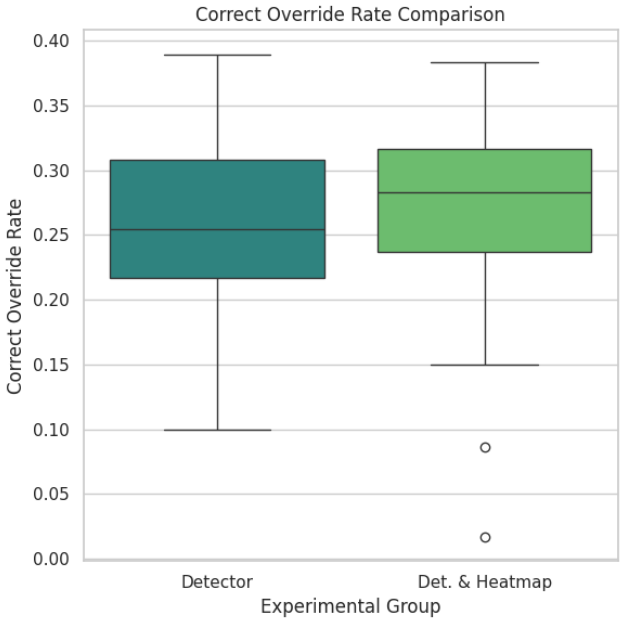


Figure 6.12: Box plot comparing the Correct Override Rate (proportion of trials where participants correctly classified an image despite an incorrect AI prediction) between the *Detector group* and *Det. & Heatmap group*.

This also follows from a closer look at Figure 6.11b, where the proportion of Agreement rate within the incorrect classification for *Det. & Heatmap group* lies below the *Detector group* curve for bins with higher but incorrect certainty (61-80, 81-100). However, in none of these cases was their statistical difference confirmed (alpha = 0.05, Mann-Whitney U test: Correct Override Rate U = 1948.50, p-value = 0.187; 61-80 bin U = 2610.0, p-value = 0.099; 81-100 bin U = 2475.5, p-value = 0.297).

Conclusion for H3: Based on the statistically significant differences found in the observed trends and the fact that the *Det. & Heatmap group* was the only one that did not show statistical significance between the recall, specificity, and precision metrics; the findings strongly confirm that heatmaps mitigate the negative impact resulting from blind trust in AI. Thanks to heatmaps, people are shown to make better, more informed, and more balanced decisions, especially in situations where the AI was wrong, which is also important in the context of the findings from H2, which shows that participants adhered to the correct prediction but in problematic cases were able to contradict the result from the detector. Hypothesis H3 is **supported** based on the data.

Conclusion to RQ2: Hypothesis H1 was rejected due to the lack of evidence of a statistically significant difference between groups based on accuracy (accuracy and F1-Score). H2 confirmed that the AI detector has a measurable impact on the decision-making process of the participant who worked with the provided score, especially in cases where the AI signaled high confidence. From H3, it follows that providing only the score from the detector is risky for incorrect AI predictions; the findings are that heatmaps effectively

mitigate dire predictions and lead to better, more informed, and more balanced decisions, and in the case of a good AI prediction, participants agree with the detector.

6.4 Analysis of Confidence & Time

This section will examine potentially interesting data from the experiment that are not directly related to the main research questions and hypotheses. The section analyzes the time spent with the experiment, its dependence on performance, and the confidence that participants chose.

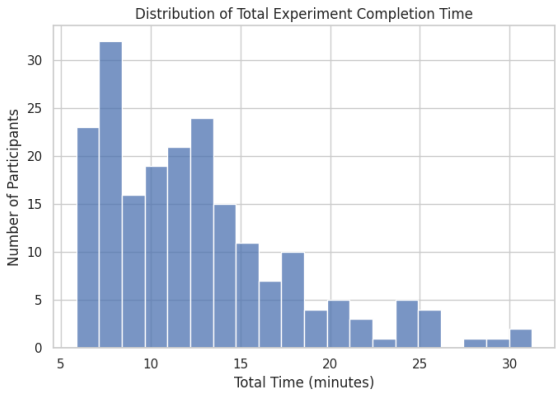
6.4.1 Time Analysis

This subsection will focus on the experiment's time side, specifically the total time spent by individual users on the entire questionnaire, including reading the introductory instructions, answering the introductory questions, classifying the images, and answering the final questions and also whether there is any dependence of time on accuracy (F1-Score) in the experiment.

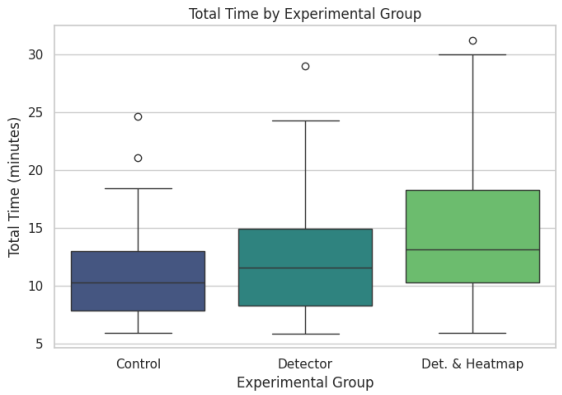
General Time Analysis

Analyzing the time side allows us to understand the time intensity between individual groups better.

The histogram in Figure 6.13a shows the overall distribution of times in seconds, resulting in a predominant representation frequency in the range of approximately 6 to 13 minutes. The median within the sample was at the level of 11 and a half minutes.



(a) Overall distribution of total time taken by participants to complete the experiment, showing a peak around 7-13 minutes.



(b) Box plots comparing total completion times across experimental groups, indicating a trend of increasing median time from *Control* to *Detector* to *Detector & Heatmap*.

Figure 6.13: Analysis of Total Experiment Completion Time (in minutes): (a) Overall distribution across all participants, and (b) comparison of completion times by experimental group.

Focusing on the comparison of time between individual groups as shown in Figure 6.13b, based on the box plots, it can be seen that there is a clear trend of a gradual increase in the median from the *Control group* through the *Detector group* to the *Det. & Heatmap group*.

To test for potential statistical significance in the observed trends between groups, the nonparametric Kruskal-Wallis test was used, which confirmed a statistically significant difference in completion time between groups (H = 11.21, p = 0.0037, alpha = 0.05).

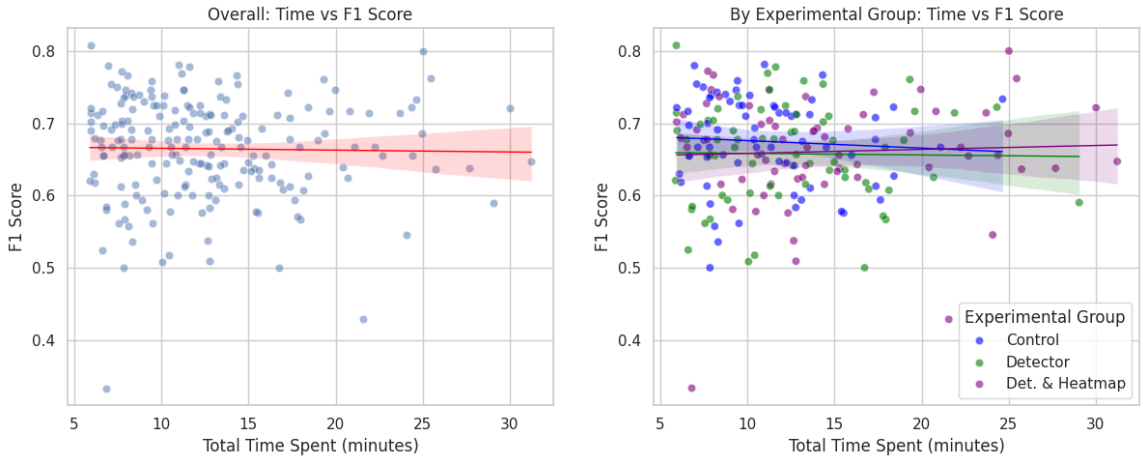
A Dunn's post-hoc test with Bonferroni correction for pairwise comparisons was then performed to determine which groups had statistically significant differences. Based on the results, a significant difference was found between the *Control group* and the *Detector group* and *Det. & Heatmap group* (p = 0.0025, alpha = 0.05). Participants assigned to *Det. & Heatmap group* took statistically significantly longer to complete the experiment than participants assigned to *Control group*. The differences between *Control group* and *Detector group* (p = 0.460, alpha = 0.05) and *Detector group* and *Det. & Heatmap group* (p = 0.175, alpha = 0.05) were not statistically significant at the alpha = 0.05 significance level.

The finding that *Det. & Heatmap group* would take longer to complete the process than the group without any additional information is in line with expectations since *Det. & Heatmap group* participants had the most information available, where the additional time probably played a role mainly in processing and interpreting heatmaps compared to *Control group*, which only classified images.

Relationship between time and F1-Score

Looking at Figure 6.14a, it is clear that across groups, time had no effect on the F1-Score; this score was almost invariant over time at the level of 0.67.

A closer look at the dependence is provided by Figure 6.14b, which shows signs of trends for individual groups. The *Control group* paradoxically achieved worse scores with increasing time, in the case of the *Detector group*, it remains the same, only the *Det. & Heatmap group* achieves slightly better results with increasing time.



(a) Scatter plot showing the overall relationship between total completion time and F1-Score, indicating no apparent correlation.

(b) Scatter plot showing the relationship between total completion time and F1-Score within each experimental group, confirming the lack of a clear correlation for any group.

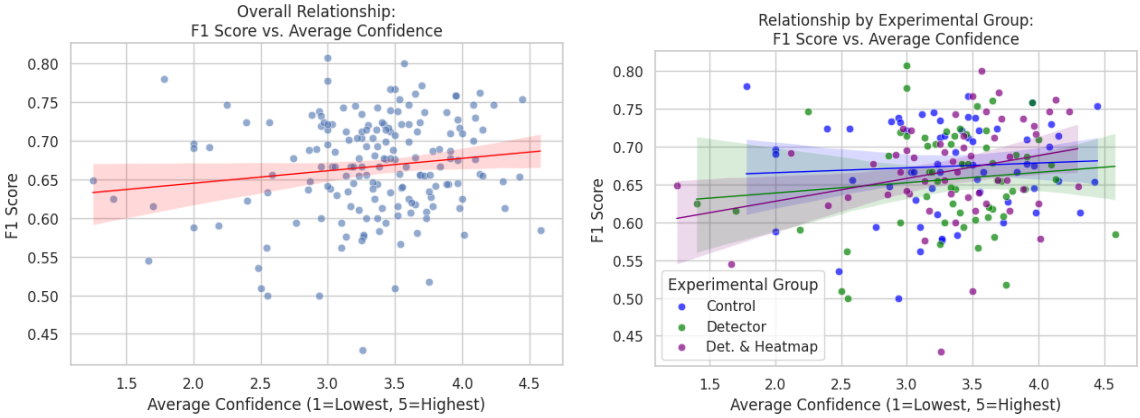
Figure 6.14: Relationship between Total Time Spent (minutes) and Participant F1-Score: (a) Overall correlation across all participants, and (b) correlation examined separately for each experimental group.

6.4.2 Confidence Analysis

This subsection will take a closer look at the impact of participants' confidence on F1-Score and the calibration of confidence between correct and incorrect trials of users.

Average Confidence vs F1-Score Analysis

Figure 6.15a shows that the higher the participants' average confidence, the better they achieved F1-Score. When broken down into smaller groups, the Figure 6.15b shows that this fact was reflected in each experimental group, although only the *Det. & Heatmap group* achieved the most considerable and statistically significant positive correlation based on Spearman's correlation with the result p = 0.0473, rho = 0.269, alpha = 0.05; the other groups achieved p values higher than 0.05. This shows that participants from *Det. & Heatmap group* could better connect their confidence in the decision with how well they performed in the classification using the heatmap.



(a) Scatter plot showing a weak positive overall correlation between average confidence and F1-Score.

(b) Scatter plot illustrating the relationship between average confidence and F1-Score within each group, showing weak positive trends, with the relationship being statistically significant only for the *Det. & Heatmap group*.

Figure 6.15: Relationship between Participant average confidence (1-5 Likert scale) and F1-Score: (a) Overall correlation across all participants, and (b) correlation examined separately for each experimental group.

Calibration of Average Confidence

Figure 6.16 compares the calibration of average confidence in correct and incorrect classified cases in individual groups; looking at the Figure, it is clear that the calibration achieved approximately the same distribution across groups and the same median values. The main finding for us is that participants across all groups were able to calibrate their confidence, which means that in correct trials, they were significantly more confident than in incorrect ones. However, the correct/incorrect medians always differed by a maximum of half a point on a Likert scale (1-5).

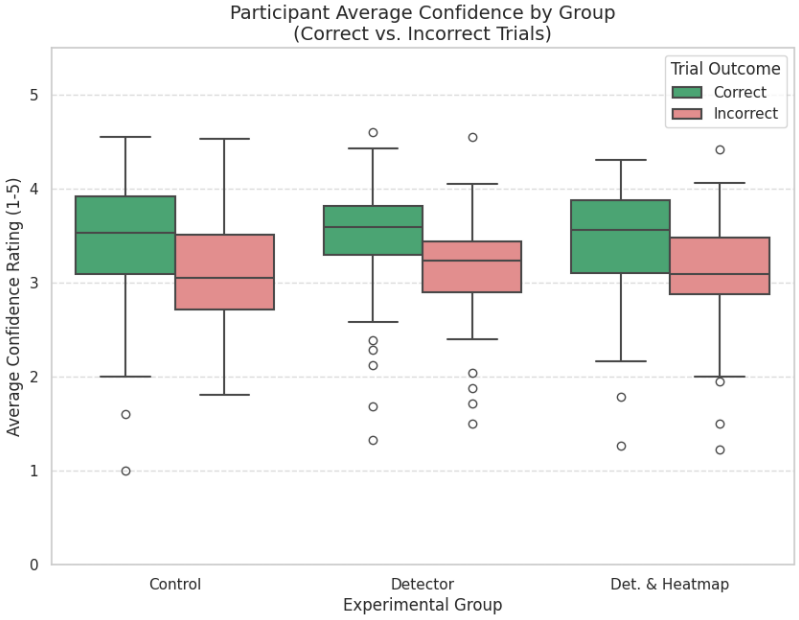


Figure 6.16: Box plots comparing participants' average confidence ratings on trials where their classification was correct versus incorrect, shown separately for each experimental group.

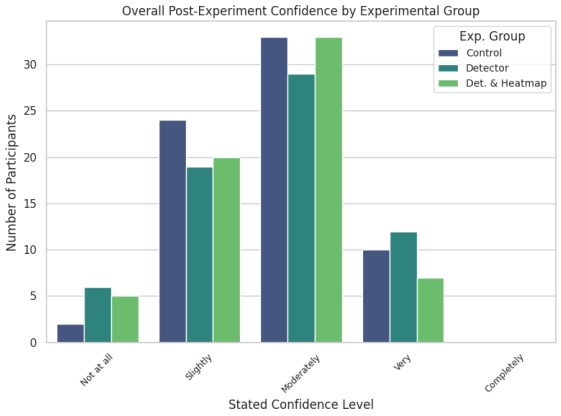
6.5 Post-Experiment Subjective Feedback

After the classification, participants were asked to answer final questions for the respective groups regarding their subjective feelings and experiences. An overview of the results is provided in this section.

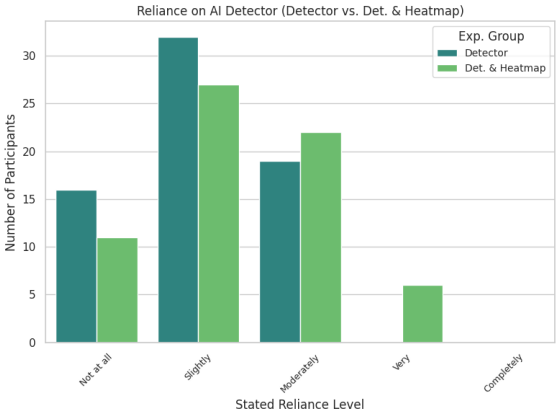
- **Overall Subjective Confidence after the Experiment:** Figure 6.17a shows that the majority of group participants chose the value 3 ("Moderately") on a Likert scale from 1 ("Not at all") to 5 ("Completely"). Fewer participants chose the other values on this scale. However, the Figure shows that people with the answer 1 ("Not at all") and 2 ("Slightly") at least prevail over people with the answer 4 ("Very"). Nobody chose option 5 ("Completely"), which means that people are slightly pessimistic about their subjective feeling about the ability to detect deepfake, which is a slight difference compared to their confidence before the start of the experiment (Figure 6.3c), where the scale was reversed.

- **Ease of understanding the information provided (*Detector group* and *Det. & Heatmap group*)**: Participants in *Detector group* and *Det. & Heatmap group* were asked how easy it was to understand the information provided. Figure 6.17b shows that most participants understood the information at the "Very" and "Completely" levels. In *Detector group*, there was a gradual increase within the scale, while in *Det. & Heatmap group*, can be observed a slight drop in the "Completely" option; some participants probably gradually moved to "Very" and "Moderately", which may be due to the heatmap, which requires a certain amount of time and skills to understand its meaning.

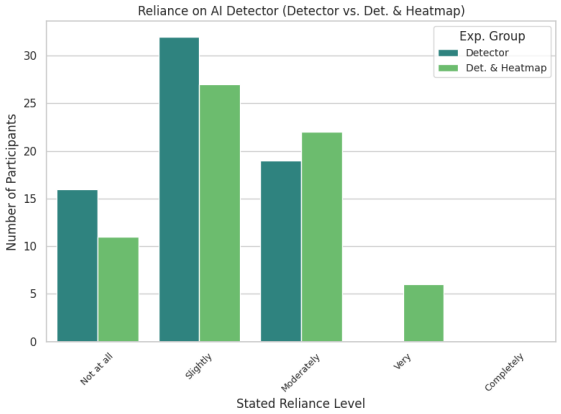
- **Reliance on the AI detector (*Detector group* and *Det. & Heatmap group*):** Another question the users were asked was their reliance on the percentage output from the detector. In Figure 6.17c, the difference between the groups can be seen; in *Detector group*, no one chose the options "Very" and "Completely", and the most frequent answer in *Detector group* and *Det. & Heatmap group* was "Slightly". In the case of *Det. & Heatmap group*, the participants' reliance on the detector was distributed within four levels. However, no one chose the option "Completely". This distribution indicates a visual deviation, which was also confirmed based on the Mann-Whitney U test, which revealed a statistically significant difference between *Detector group* and *Det. & Heatmap group* with the result (U = 1803.0, p = 0.0499, alpha = 0.05). A possible explanation for the difference is that the participants relied on the output from the detector more because they had another important context to look at, which helped them increase their confidence. The reliance itself was qualitatively different and better substantiated, which also supports the conclusions of Hypothesis 3.
- **Perceived usefulness of heatmaps (*Det. & Heatmap group* only):** Participants in *Det. & Heatmap group* were asked about their subjective feelings about the degree of usefulness of the presented heatmap. The results from Figure 6.17d are that the majority of participants chose the options "Not at all", "Slightly" (most common), and "Moderately", which is surprising because, based on Hypothesis 3, were found that heatmaps have a significant positive impact on the classification process. This led to a contradiction between the objective findings, based on detailed analysis, and the subjective feeling of the usefulness of heatmaps. The question thus arises of whether heatmaps are the best possible XAI technology, which will be discussed later.



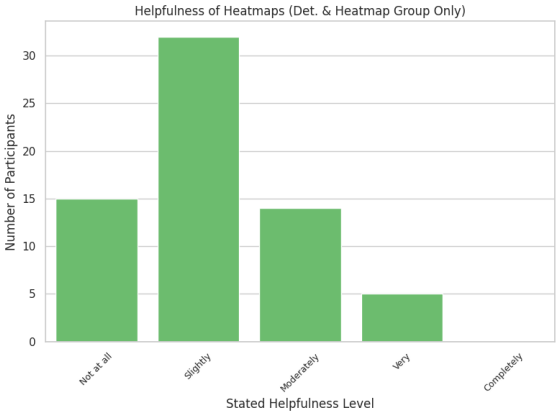
(a) Distribution of participants' self-reported overall confidence level after the experiment, showing similar distributions peaking at 'Moderately' across all groups.



(b) Participants' ratings of the ease of understanding the provided AI assistance, indicating high perceived ease for both the *Detector group* and *Det. & Heatmap group*.



(c) Participants' self-reported reliance on the AI detector, showing significantly higher reliance reported by the *Det. & Heatmap group* compared to the *Detector group*.



(d) Participants' ratings of the helpfulness of heatmaps (*Det. & Heatmap group* only), with the most frequent response being 'Slightly' helpful.

Figure 6.17: Participant subjective ratings from the post-experiment questionnaire: Overall confidence in classifications by group, Ease of understanding AI assistance (*Detector group* vs. *Det. & Heatmap group*), Reliance on AI detector (*Detector group* vs. *Det. & Heatmap group*), and Perceived helpfulness of heatmaps (*Det. & Heatmap group* only).

6.6 Discussion

This section summarizes all findings and describes their implications for designing and using AI-assisted deepfake detection systems, describes the study's limitations, and provides potential suggestions for future research.

6.6.1 Summary and Interpretation of Key Findings

The experiment provides a comprehensive overview of the issues of deepfakes detection and assisted decision-making using XAI, with the following summary.

Basic human ability and the impact of AI on overall performance. The answers to this can be found in the analysis and conclusions for RQ1 and H1. They show that the level of human ability to detect deepfakes in the context of a specifically set experiment containing challenging and ambiguous cases is 67%, and this value is the same across all groups (rejected H1). However, the conclusion from the results was that the participants in the study were biased by the knowledge that they were going to classify images that also contained deepfakes. Thanks to this, they classified a larger number of images as deepfakes, which increased the accuracy of correct classification of deepfakes, but on the contrary, increased the False Positive metric. However, AI assistance was very beneficial in cases where its prediction was correct, which only supports the claim that the more powerful the detector we have, the better results we will achieve with the help of AI compared to cases where we would not have this help.

Strong influence of AI score and its risks. AI has a fundamental influence on the decision-making process for humans, especially in cases where the detector demonstrates a prediction with high confidence (RQ2). In cases where the detector was wrong, the user agreement rate was at a level of around 50%, which indicates that participants did not just blindly follow the detector's prediction. However, they could correct this error on every second classified image (H2 supported) in these cases.

Mitigation role of heatmap. The heatmap fulfilled its function, and people made better and more informed decisions, especially in the case of incorrect AI predictions, because in those cases, the heatmap should be an additional factor in the decision-making process. It was observed that, with incorrect predictions, what was the subject of H2, in the case of the *Detector group*, the success rate decreased statistically significantly compared to the *Control group*; on the contrary, in the *Det. & Heatmap group*: The heatmap was able to mitigate this error. The success rate of this group did not exceed that of the *Control group*, but it is important that it was not worse. Another finding that supports the claim is that participants in the heatmap group achieved statistically better F1-Score depending on the time they devoted to classification. Heatmaps thus help mitigate blind trust in the percentage output from the detector (H3 supported).

This demonstrated the potential of XAI methods, such as heatmaps, which provide insight into the detector's reasoning and could be particularly important in a professional context. For example, forensic experts, such as police analysts, need tools that provide predictions and explain why a particular conclusion was reached; reliably interpreted heatmaps could fulfill this need and support investigations.

Subjective perception vs. objective impact of heatmaps. A paradoxical finding was that the subjective user perception of the helpfulness of the heatmap within *Det. & Heatmap group* did not correspond to the objectively proven impacts. However, they also relied more on the correct percentage score from the detector statistically significantly due

to the additional context that the heatmap provides. The result of this finding may be that despite the objective positive impact, the heatmap (or selected technique Grad-CAM++) may not be the ideal approach for providing additional context.

6.6.2 Limitations

The study certainly has limitations that need to be considered.

- **Participant Sample:** The participant sample was large, but based on demographic data, it is clear that it was not balanced, with a dominant representation of young male students with a technical education.
- **Experimental Environment:** An online questionnaire, in which participants were informed beforehand about what would happen, may not fully reflect the complexity and pressure of real-world situations.
- **Materials:** The training, testing, and validation datasets of facial images were created using only videos from the FaceForensics++ dataset.
- **XAI method:** Focusing exclusively on one option that XAI offers (heatmap) and also only one technology for its creation (Grad-CAM++).
- **Task design:** a specific setup with 60 images, uniform score distribution, and 50% AI error rate was designed to test "gray zones", but typically it may not match the performance of current state-of-the-art solutions.

6.6.3 Future Work

Based on the findings, the following suggestions for future research were proposed.

- **Comparison and design of XAI methods:** To compare other technologies for creating heatmaps described in this work (LIME, SHAP, RISE) or to choose a completely different XAI method (text explanations or other visualization approaches), which will be technically correct, efficient and at the same time easily interpretable for a wide target group.
- **Impact of training and literacy:** To determine whether targeted training focused on interpreting XAI outputs can improve the ability of users to use AI assistance and its explanations more effectively.
- **Testing in more realistic conditions:** Verify the findings with other types of deepfakes (audio, video) in scenarios closer to reality (e.g., social networks, news).
- **Sample expansion and confirmation of conclusions:** To achieve a more balanced sample regarding demographics, education, and experience and subsequently confirm the conclusions.

Chapter 7

Conclusion

With the increasing use of deepfake technology, the need for explainable detection systems is more important than ever. This thesis emphasizes the importance of integrating explainability mechanisms into black-box models to increase their transparency and usability. More specifically, this work investigated the impact of deepfake detection tools and their visualizations on the ability of users to identify deepfakes correctly and on their confidence in their decisions. Experimental findings indicate a complex interplay between AI assistance, human judgment, and the way information is presented.

The main element of the thesis was the proposed experiment involving three groups of participants: a *Control group*, a *Detector group* with deepfake detector outputs at its disposal, and a *Det. & Heatmap group* that will be assisted in its decision-making by heatmap and the percentage output from the deepfake detector. To prepare for the experiment, it was necessary to study various deepfake detection models and their explainability methods; the models were located in publicly available repositories that contain potential candidates for use in the experimental work. Upon closer examination, it was found that the Grad-CAM++ technology for generating heatmaps prevails among the others in the context of deepfake detection. The experiment evaluated how the additional information affected the participant's ability to distinguish genuine images from deepfakes correctly, specifically focusing on challenging cases designed to probe the limits of human-AI collaboration.

The results revealed that baseline human performance in this challenging task was around 67% (F1-score), significantly better than chance but exhibiting a bias towards identifying images as fake (RQ1). Overall, providing AI assistance (*Detector group* and *Det. & Heatmap group*) did not lead to a statistically significant improvement in the F1-score compared to the *Control group* in this experimental setup (H1 rejected). However, assistance was beneficial when the AI prediction was correct. Crucially, the AI detector's score strongly influenced user decisions (H2 supported). However, this reliance proved detrimental when the AI was incorrect, leading to significantly worse performance in the *Detector group* compared to the *Control group*. Adding heatmap effectively mitigated this negative impact for *Det. & Heatmap group*, keeping performance on par with the *Control group* when the AI erred, significantly improving the confidence-weighted score (CWS) compared to *Detector group* in those situations and leading to a more balanced decision-making approach (H3 supported).

Although Grad-CAM++ heatmap based on the results of the studies did not dramatically increase the accuracy of identifying detector errors, the findings demonstrate its important role in mitigating the risks of blind trust in potentially misleading detector scores and also allowed participants to calibrate their confidence better. However, the paradox-

ical finding is that despite objectively better results, participants in the *Det. & Heatmap group* subjectively rated heatmaps as unhelpful or only moderately helpful while reporting a significantly higher level of reliance on the combined system compared to the *Detector group*.

This finding only confirms a critical challenge for XAI, where these mechanisms demonstrate a better and more robust human-AI decision-making process; their current form (heatmap and Grad-CAM++) may not be intuitive or interpretable enough for users to appreciate or consciously exploit their benefits entirely.

As society grapples with the implications of generative AI, fostering trust and accessibility through explainable systems will be paramount in protecting information integrity and security. Addressing the aforementioned future challenges associated with explainability, including exploring more user-friendly XAI methods, considering the impact of training, and ensuring scalability, will be key to achieving robust, transparent, and, most importantly, usable deepfakes detection tools for a wide range of users, and with the increasing quality of deepfakes being created, also a necessary step towards building resilience in the increasingly complex digital environment.

Bibliography

[1] *What is a neural network?* 2024. Available at: https://www.ibm.com/think/topics/neural-networks.

[2] Abir, W. H.; Khanam, F. R.; Alam, K. N.; Hadjouni, M.; Elmannai, H. et al. Detecting deepfake images using deep learning techniques and explainable AI methods. *Intelligent Automation & Soft Computing*, 2023, vol. 35, no. 2, p. 2151–2169. Available at: https://cdn.techscience.cn/ueditor/files/iasc/TSP_IASC-35-2/TSP_IASC_29653/TSP_IASC_29653.pdf.

[3] Agarwal, M.; Mukhopadhyay, R.; Namboodiri, V. P. and Jawahar, C. V. Audio-Visual Face Reenactment. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. January 2023, p. 5178–5187. Available at: https://openaccess.thecvf.com/content/WACV2023/html/Agarwal_Audio-Visual_Face_Reenactment_WACV_2023_paper.html.

[4] Almars, A. M. Deepfakes detection techniques using deep learning: a survey. *Journal of Computer and Communications*, 2021, vol. 9, no. 05, p. 20–35. Available at: https://www.scirp.org/journal/paperinformation?paperid=109149.

[5] Aouf, R. S. *Museum creates deepfake Salvador Dalí to greet visitors*. 2019. Available at: https://www.dezeen.com/2019/05/24/salvador-dali-deepfake-dali-musuem-florida/.

[6] Arya, M.; Goyal, U.; Chawla, S. et al. A Study on Deep Fake Face Detection Techniques. In: IEEE. *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. 2024, p. 459–466. Available at: https://ieeexplore.ieee.org/document/10575149.

[7] Bandara, P. *FBI Warns Deepfakes Might Be Used in Remote Job Interviews*. 2022. Available at: https://petapixel.com/2022/07/05/fbi-warns-deepfakes-might-be-used-in-remote-job-interviews/.

[8] Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S. et al. Beyond accuracy: The role of mental models in human-AI team performance. In: *Proceedings of the AAAI conference on human computation and crowdsourcing*. 2019, vol. 7, p. 2–11. Available at: https://ojs.aaai.org/index.php/HCOMP/article/view/5285.

[9] Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B. et al. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, p. 1–16. Available at: https://dl.acm.org/doi/abs/10.1145/3411764.3445717.

[10] Bhattacharyya, C.; Wang, H.; Zhang, F.; Kim, S. and Zhu, X. Diffusion deepfake. *ArXiv preprint arXiv:2404.01579*, 2024. Available at: https://arxiv.org/abs/2404.01579.

[11] Bitouk, D.; Kumar, N.; Dhillon, S.; Belhumeur, P. and Nayar, S. K. Face swapping: automatically replacing faces in photographs. In: *ACM SIGGRAPH 2008 papers*. 2008, p. 1–8. Available at: https://dl.acm.org/doi/pdf/10.1145/1399504.1360638.

[12] Blanz, V.; Scherbaum, K.; Vetter, T. and Seidel, H.-P. Exchanging faces in images. In: Wiley Online Library. *Computer Graphics Forum*. 2004, vol. 23, no. 3, p. 669–676. Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-8659.2004.00799.x.

[13] Bonaccio, S. and Dalal, R. S. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*. Elsevier, 2006, vol. 101, no. 2, p. 127–151. Available at: https://www.sciencedirect.com/science/article/pii/S0749597806000719.

[14] Borade, S.; Jain, N.; Patel, B.; Kumar, V.; Godhrawala, M. et al. ResNet50 DeepFake Detector: Unmasking Reality. *Indian Journal of Science and Technology*, 2024, vol. 17, no. 13, p. 1263–1271. Available at: https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2024/Issue-13/IJST-2024-285.pdf.

[15] Bounareli, S.; Argyriou, V. and Tzimiropoulos, G. Finding directions in gan's latent space for neural face reenactment. *ArXiv preprint arXiv:2202.00046*, 2022. Available at: https://arxiv.org/abs/2202.00046.

[16] Bounareli, S.; Tzelepis, C.; Argyriou, V.; Patras, I. and Tzimiropoulos, G. Stylemask: Disentangling the style space of stylegan2 for neural face reenactment. In: IEEE. *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)*. 2023, p. 1–8. Available at: https://ieeexplore.ieee.org/abstract/document/10042744.

[17] Brooks, T.; Holynski, A. and Efros, A. A. InstructPix2Pix: Learning to Follow Image Editing Instructions. In: IEEE. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, p. 18392–18402. Available at: https://ieeexplore.ieee.org/abstract/document/10204579.

[18] Burton, J. W.; Stein, M.-K. and Jensen, T. B. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making*. Wiley Online Library, 2020, vol. 33, no. 2, p. 220–239. Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1002/bdm.2155.

[19] Channel4. *Deepfake queen to deliver Channel 4 Christmas message*. 2020. Available at: https://www.bbc.com/news/technology-55424730.

[20] Chattopadhay, A.; Sarkar, A.; Howlader, P. and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE. *2018 IEEE winter conference on applications of*

computer vision (WACV). 2018, p. 839–847. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8354201.

[21] Chen, L.; Chen, J.; Hajimirsadeghi, H. and Mori, G. Adapting grad-cam for embedding networks. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020, p. 2794–2803. Available at: https://openaccess.thecvf.com/content_WACV_2020/html/Chen_Adapting_Grad-CAM_for_Embedding_Networks_WACV_2020_paper.html.

[22] Chen, L.; Maddox, R. K.; Duan, Z. and Xu, C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, p. 7832–7841. Available at: https://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Hierarchical_Cross-Modal_Talking_Face_Generation_With_Dynamic_Pixel-Wise_Loss_CVPR_2019_paper.html.

[23] Chen, T.; Kornblith, S.; Norouzi, M. and Hinton, G. A simple framework for contrastive learning of visual representations. In: PMLR. *International conference on machine learning*. 2020, p. 1597–1607. Available at: https://proceedings.mlr.press/v119/chen20j.html.

[24] Chen, Y.; Haldar, N. A. H.; Akhtar, N. and Mian, A. Text-image guided Diffusion Model for generating Deepfake celebrity interactions. In: IEEE. *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2023, p. 348–355. Available at: https://ieeexplore.ieee.org/abstract/document/10410943.

[25] Choo, J. and Liu, S. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*. IEEE, 2018, vol. 38, no. 4, p. 84–92. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8402187.

[26] Cicman, D. *Deepfake video call with me*. 2023. Available at: https://www.linkedin.com/posts/dalibor-cicman_deepfake-videohovor-so-mnou-tento-activity-7096111558144995328-_7zK/.

[27] Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B. et al. Generative adversarial networks: An overview. *IEEE signal processing magazine*. IEEE, 2018, vol. 35, no. 1, p. 53–65. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8253599.

[28] Dang, H.; Liu, F.; Stehouwer, J.; Liu, X. and Jain, A. K. On the detection of digital face manipulation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*. 2020, p. 5781–5790. Available at: https://openaccess.thecvf.com/content_CVPR_2020/html/Dang_On_the_Detection_of_Digital_Face_Manipulation_CVPR_2020_paper.html.

[29] Dang, M. and Nguyen, T. N. Digital face manipulation creation and detection: A systematic review. *Electronics*. MDPI, 2023, vol. 12, no. 16, p. 3407. Available at: https://www.mdpi.com/2079-9292/12/16/3407.

[30] Davies, G. *David Beckham 'speaks' 9 languages for new campaign to end malaria*. 2019. Available at: https://abcnews.go.com/International/david-beckham-speaks-languages-campaign-end-malaria/story?id=62270227.

[31] Deb, D.; Zhang, J. and Jain, A. K. Advfaces: Adversarial face synthesis. In: IEEE. *2020 IEEE International Joint Conference on Biometrics (IJCB)*. 2020, p. 1–10. Available at: https://ieeexplore.ieee.org/abstract/document/9304898.

[32] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K. et al. Imagenet: A large-scale hierarchical image database. In: Ieee. *2009 IEEE conference on computer vision and pattern recognition*. 2009, p. 248–255. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5206848.

[33] Dhanyalakshmi, R.; Popirlan, C.-I. and Hemanth, D. J. A survey on deep learning based reenactment methods for deepfake applications. *IET Image Processing*. Wiley Online Library, 2024. Available at: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/ipr2.13201.

[34] Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 2021, vol. 34, p. 8780–8794. Available at: https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

[35] Dietvorst, B. J.; Simmons, J. P. and Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*. American Psychological Association, 2015, vol. 144, no. 1, p. 114. Available at: https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf.

[36] Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H. et al. Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In: IEEE. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, p. 3994–4004. Available at: https://ieeexplore.ieee.org/abstract/document/10204669.

[37] Farhood, H.; Najafi, M. and Saberi, M. Improving Deep Learning Transparency: Leveraging the Power of LIME Heatmap. In: Springer. *International Conference on Service-Oriented Computing*. 2023, p. 72–83. Available at: https://link.springer.com/chapter/10.1007/978-981-97-0989-2_7.

[38] Fel, T.; Cadène, R.; Chalvidal, M.; Cord, M.; Vigouroux, D. et al. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in neural information processing systems*, 2021, vol. 34, p. 26005–26014. Available at: https://proceedings.neurips.cc/paper/2021/hash/da94cbeff56cfda50785df477941308b-Abstract.html.

[39] Fick, G. and Sprague, R. H. Decision Support Systems: Issues and Challenges: Proceedings of an International Task Force Meeting June 23-25, 1980. Elsevier, 2013.

[40] Firc, A. Applicability of Deepfakes in the Field of Cyber Security. *Brno University of Technology, Faculty of Information Technology, Brno. Supervisor Mgr. Kamil Malinka, Ph. D*, 2021. Available at: https://theses.cz/id/dwy6w2/23761.pdf.

[41] Firc, A.; Malinka, K. and Hanáček, P. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*. Elsevier, 2023,

vol. 9, no. 4. Available at:
https://www.cell.com/heliyon/fulltext/S2405-8440(23)02297-1.

[42] Fosco, C. L. *Detecting Deepfakes with Human Help to Help Humans Detect Deepfakes*. 2023. Dissertation. Massachusetts Institute of Technology. Available at: https://dspace.mit.edu/handle/1721.1/154206.

[43] Galdi, C.; Panariello, M.; Todisco, M. and Evans, N. 2D-Malafide: Adversarial Attacks Against Face Deepfake Detection Systems. In: IEEE. *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2024, p. 1–7. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10786754.

[44] Gongane, V. U.; Munot, M. V. and Anuse, A. D. A survey of explainable ai techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*. Springer, 2024, p. 1–37. Available at: https://link.springer.com/article/10.1007/s42001-024-00248-9.

[45] Gowrisankar, B. and Thing, V. L. An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. *Computers & Security*. Elsevier, 2024, vol. 139, p. 103684. Available at: https://www.sciencedirect.com/science/article/pii/S0167404823005941.

[46] Green, B. and Chen, Y. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*. ACM New York, NY, USA, 2019, vol. 3, CSCW, p. 1–24. Available at: https://dl.acm.org/doi/abs/10.1145/3359152.

[47] Grgić Hlača, N.; Engel, C. and Gummadi, K. P. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on human-computer interaction*. ACM New York, NY, USA, 2019, vol. 3, CSCW, p. 1–25. Available at: https://dl.acm.org/doi/abs/10.1145/3359280.

[48] Groh, M. *Detect DeepFakes: How to counteract misinformation created by AI*. 2020. Available at: https://www.media.mit.edu/projects/detect-fakes/overview/.

[49] Groshev, A.; Maltseva, A.; Chesakov, D.; Kuznetsov, A. and Dimitrov, D. GHOST—a new face swap approach for image and video domains. *IEEE Access*. IEEE, 2022, vol. 10, p. 83452–83462. Available at: https://ieeexplore.ieee.org/abstract/document/9851423.

[50] Highton, J.; Chong, Q. Z.; Crawley, R.; Schnabel, J. A. and Bhatia, K. K. Evaluation of Randomized Input Sampling for Explanation (RISE) for 3D XAI-Proof of Concept for Black-Box Brain-Hemorrhage Classification. In: Springer. *International Conference on Medical Imaging and Computer-Aided Diagnosis*. 2023, p. 41–51. Available at: https://link.springer.com/chapter/10.1007/978-981-97-1335-6_4.

[51] Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R. et al. Imagen video: High definition video generation with diffusion models. *ArXiv preprint arXiv:2210.02303*, 2022. Available at: https://arxiv.org/abs/2210.02303.

[52] Ho, J.; Jain, A. and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020, vol. 33, p. 6840–6851. Available at: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[53] Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q. et al. Implicit identity driven deepfake face swapping detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, p. 4490–4499. Available at: https://ieeexplore.ieee.org/abstract/document/10204862.

[54] IBM. *What is machine learning?* 2024. Available at: https://www.ibm.com/think/topics/machine-learning.

[55] Johansen, A. G. *How to spot deepfake videos — 15 signs to watch for*. 2020. Available at: https://us.norton.com/internetsecurity-emerging-threats-how-to-spot-deepfakes.html.

[56] Karn, A.; Kumar, S.; Kushwaha, S. K. and Katarya, R. Image Synthesis Using GANs and Diffusion Models. In: IEEE. *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*. 2023, vol. 1, p. 1–6. Available at: https://ieeexplore.ieee.org/abstract/document/10263208.

[57] Karras, T. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *ArXiv preprint arXiv:1710.10196*, 2017. Available at: https://arxiv.org/abs/1710.10196.

[58] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J. et al. Analyzing and Improving the Image Quality of StyleGAN. In: IEEE. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, p. 8107–8116. Available at: https://ieeexplore.ieee.org/abstract/document/9156570.

[59] Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H. et al. Imagic: Text-based real image editing with diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, p. 6007–6017. Available at: https://openaccess.thecvf.com/content/CVPR2023/html/Kawar_Imagic_Text-Based_Real_Image_Editing_With_Diffusion_Models_CVPR_2023_paper.html.

[60] Kietzmann, J.; Lee, L. W.; McCarthy, I. P. and Kietzmann, T. C. Deepfakes: Trick or treat? *Business Horizons*. Elsevier, 2020, vol. 63, no. 2, p. 135–146. Available at: https://www.sciencedirect.com/science/article/pii/S0007681319301600.

[61] Kim, H.; Choi, Y.; Kim, J.; Yoo, S. and Uh, Y. Exploiting spatial dimensions of latent in gan for real-time image editing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, p. 852–861. Available at: https://openaccess.thecvf.com/content/CVPR2021/html/Kim_Exploiting_Spatial_Dimensions_of_Latent_in_GAN_for_Real-Time_Image_CVPR_2021_paper.html.

[62] Korshunov, P. and Marcel, S. Deepfake detection: humans vs. machines. *ArXiv preprint arXiv:2009.03155*, 2020. Available at: https://arxiv.org/abs/2009.03155.

[63] Kwak, J.-g.; Li, Y.; Yoon, D.; Kim, D.; Han, D. et al. Injecting 3d perception of controllable nerf-gan into stylegan for editable portrait image synthesis. In: Springer. *European Conference on Computer Vision*. 2022, p. 236–253. Available at: https://link.springer.com/chapter/10.1007/978-3-031-19790-1_15.

[64] Lad, S. Applied Ethical and Explainable AI in Adversarial Deepfake Detection: From Theory to Real-World Systems. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 2024, vol. 6, no. 1, p. 126–137. Available at: https://ojs.boulibrary.com/index.php/JAIGS/article/view/236.

[65] LeCun, Y.; Bengio, Y. and Hinton, G. Deep learning. *Nature*. Nature Publishing Group UK London, 2015, vol. 521, no. 7553, p. 436–444. Available at: https://www.nature.com/articles/nature14539.

[66] Lee, H.; Lee, C.; Farhat, K.; Qiu, L.; Geluso, S. et al. The Tug-of-War Between Deepfake Generation and Detection. *ArXiv preprint arXiv:2407.06174*, 2024. Available at: https://arxiv.org/abs/2407.06174.

[67] Lewandowsky, S.; Ecker, U. K.; Seifert, C. M.; Schwarz, N. and Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*. Sage Publications Sage CA: Los Angeles, CA, 2012, vol. 13, no. 3, p. 106–131. Available at: https://journals.sagepub.com/doi/full/10.1177/1529100612451018.

[68] Li, L.; Bao, J.; Yang, H.; Chen, D. and Wen, F. Faceshifter: Towards high fidelity and occlusion aware face swapping. *ArXiv preprint arXiv:1912.13457*, 2019. Available at: https://arxiv.org/abs/1912.13457.

[69] Li, Q.; Wang, W.; Xu, C.; Sun, Z. and Yang, M.-H. Learning disentangled representation for one-shot progressive face swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2024. Available at: https://ieeexplore.ieee.org/abstract/document/10536627.

[70] Malik, A.; Kuribayashi, M.; Abdullahi, S. M. and Khan, A. N. DeepFake detection for human face images and videos: A survey. *Ieee Access*. IEEE, 2022, vol. 10, p. 18757–18775. Available at: https://ieeexplore.ieee.org/abstract/document/9712265.

[71] Marra, F.; Gragnaniello, D.; Verdoliva, L. and Poggi, G. Do gans leave artificial fingerprints? In: IEEE. *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*. 2019, p. 506–511. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8695364.

[72] Matern, F.; Riess, C. and Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE. *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 2019, p. 83–92. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8638330.

[73] Metz, R. *How a deepfake Tom Cruise on TikTok turned into a very real AI company*. 2021. Available at: https://edition.cnn.com/2021/08/06/tech/tom-cruise-deepfake-tiktok-company/index.html.

[74] Mirsky, Y. and Lee, W. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*. ACM New York, NY, USA, 2021, vol. 54, no. 1, p. 1–41. Available at: https://dl.acm.org/doi/pdf/10.1145/3425780.

[75] Naitali, A.; Ridouani, M.; Salahdine, F. and Kaabouch, N. Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Computers*. MDPI, 2023, vol. 12, no. 10, p. 216. Available at: https://www.mdpi.com/2073-431X/12/10/216.

[76] Naruniec, J.; Helminger, L.; Schroers, C. and Weber, R. M. High-resolution neural face swapping for visual effects. In: Wiley Online Library. *Computer Graphics Forum*. 2020, vol. 39, no. 4, p. 173–184. Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.14062.

[77] Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P. et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In: PMLR. *International Conference on Machine Learning*. 2022, p. 16784–16804. Available at: https://proceedings.mlr.press/v162/nichol22a.html.

[78] Park, J.; Kang, H. and Kim, H. Y. Human, do you think this painting is the work of a real artist? *International Journal of Human–Computer Interaction*. Taylor & Francis, 2024, vol. 40, no. 18, p. 5174–5191. Available at: https://www.tandfonline.com/doi/epdf/10.1080/10447318.2023.2232978.

[79] Pashine, S.; Mandiya, S.; Gupta, P. and Sheikh, R. Deep fake detection: survey of facial manipulation detection solutions. *ArXiv preprint arXiv:2106.12605*, 2021. Available at: https://arxiv.org/abs/2106.12605.

[80] Patel, Y.; Tanwar, S.; Bhattacharya, P.; Gupta, R.; Alsuwian, T. et al. An improved dense CNN architecture for deepfake image detection. *IEEE Access*. IEEE, 2023, vol. 11, p. 22081–22095. Available at: https://ieeexplore.ieee.org/abstract/document/10057390.

[81] Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S. et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *ArXiv preprint arXiv:2005.05535*, 2020. Available at: https://arxiv.org/abs/2005.05535.

[82] Petsiuk, V. Rise: Randomized Input Sampling for Explanation of black-box models. *ArXiv preprint arXiv:1806.07421*, 2018. Available at: https://arxiv.org/abs/1806.07421.

[83] Pinkney, J. N. and Li, C. Clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip. *ArXiv preprint arXiv:2210.02347*, 2022. Available at: https://arxiv.org/abs/2210.02347.

[84] Rangarajan, P. K.; Sukesh, M.; Abinandhini, D.; Jaikanth, Y. et al. Detecting AI-generated images with CNN and Interpretation using Explainable AI. In: IEEE. *2024 IEEE International Conference on Contemporary Computing and*

Communications (InC4). 2024, vol. 1, p. 1–6. Available at: https://ieeexplore.ieee.org/abstract/document/10649158.

[85] Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K. R.; Dhurandhar, A. et al. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*. ACM New York, NY, USA, 2022, vol. 6, CSCW1, p. 1–22. Available at: https://dl.acm.org/doi/abs/10.1145/3512930.

[86] Rehaan, M.; Kaur, N. and Kingra, S. Face manipulated deepfake generation and recognition approaches: A survey. *Smart Science*. Taylor & Francis, 2024, vol. 12, no. 1, p. 53–73. Available at: https://www.tandfonline.com/doi/full/10.1080/23080477.2023.2268380.

[87] Remya Revi, K.; Vidya, K. and Wilscy, M. Detection of deepfake images created using generative adversarial networks: A review. In: Springer. *Second International Conference on Networks and Advances in Computational Technologies: NetACT 19*. 2021, p. 25–35. Available at: https://link.springer.com/chapter/10.1007/978-3-030-49500-8_3.

[88] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P. and Ommer, B. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, p. 10684–10695. Available at: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

[89] Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J. et al. FaceForensics++: Learning to Detect Manipulated Facial Images. In: IEEE. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, p. 1–11. Available at: https://ieeexplore.ieee.org/abstract/document/9010912.

[90] Sabae, M. S.; Dardir, M. A.; Eskarous, R. T. and Ebbed, M. R. Stylet2f: Generating human faces from textual description using stylegan2. *ArXiv preprint arXiv:2204.07924*, 2022. Available at: https://arxiv.org/abs/2204.07924.

[91] Sainburg, T.; Thielk, M.; Theilman, B.; Migliori, B. and Gentner, T. Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. *ArXiv preprint arXiv:1807.06650*, 2018. Available at: https://arxiv.org/abs/1807.06650.

[92] Scott, M.; Su In, L. et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. Curran Associates, Inc, 2017, vol. 30, p. 4765–4774. Available at: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[93] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. 2017, p. 618–626. Available at: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

[94] Shen, Y.; Yang, C.; Tang, X. and Zhou, B. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*. IEEE, 2020, vol. 44, no. 4, p. 2004–2018. Available at: https://ieeexplore.ieee.org/abstract/document/9241434.

[95] Shiohara, K. and Yamasaki, T. Detecting deepfakes with self-blended images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, p. 18720–18729. Available at: https://openaccess.thecvf.com/content/CVPR2022/papers/Shiohara_Detecting_Deepfakes_With_Self-Blended_Images_CVPR_2022_paper.pdf.

[96] Silva, S. H.; Bethany, M.; Votto, A. M.; Scarff, I. H.; Beebe, N. et al. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*. Elsevier, 2022, vol. 4, p. 100217. Available at: https://www.sciencedirect.com/science/article/pii/S2589871X2200002X.

[97] Soudy, A. H.; Sayed, O.; Tag Elser, H.; Ragab, R.; Mohsen, S. et al. Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications*. Springer, 2024, vol. 36, no. 31, p. 19759–19775. Available at: https://link.springer.com/article/10.1007/s00521-024-10181-7.

[98] Steyvers, M. and Kumar, A. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*. Sage Publications Sage CA: Los Angeles, CA, 2024, vol. 19, no. 5, p. 722–734. Available at: https://journals.sagepub.com/doi/full/10.1177/17456916231181102.

[99] Steyvers, M.; Tejeda, H.; Kerrigan, G. and Smyth, P. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*. National Acad Sciences, 2022, vol. 119, no. 11, p. e2111547119. Available at: https://www.pnas.org/doi/abs/10.1073/pnas.2111547119.

[100] Stryker, C. and Kavlakoglu, E. *What is artificial intelligence (AI)?* 2024. Available at: https://www.ibm.com/think/topics/artificial-intelligence.

[101] Sultan, D. A. and Ibrahim, L. M. Deepfake Detection Model Based on VGGFace with Head Pose Estimation Technique. In: Springer. *National Conference on New Trends in Information and Communications Technology Applications*. 2023, p. 106–117. Available at: https://link.springer.com/chapter/10.1007/978-3-031-62814-6_8.

[102] Taeb, M. and Chi, H. Comparison of deepfake detection techniques through deep learning. *Journal of Cybersecurity and Privacy*. MDPI, 2022, vol. 2, no. 1, p. 89–106. Available at: https://www.mdpi.com/2624-800X/2/1/7.

[103] Tahir, R.; Batool, B.; Jamshed, H.; Jameel, M.; Anwar, M. et al. Seeing is believing: Exploring perceptual differences in deepfake videos. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, p. 1–16. Available at: https://dl.acm.org/doi/abs/10.1145/3411764.3445699.

[104] Tan, M. and Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Chaudhuri, K. and Salakhutdinov, R., ed. *Proceedings of*

the 36th International Conference on Machine Learning. PMLR, 09–15 Jun 2019, vol. 97, p. 6105–6114. Proceedings of Machine Learning Research. Available at: https://proceedings.mlr.press/v97/tan19a.html.

[105] Terning, J. *Deep Fake of Barack Obama*. 2021. Available at: https://video.ucdavis.edu/media/Deep+Fake+of+Barack+Obama/1_6zmvebuf.

[106] Thaler, S. *AI-generated nude images of girls at NJ high school trigger police probe: 'I am terrified'*. 2023. Available at: https://nypost.com/2023/11/02/news/ai-generated-nudes-of-girls-at-nj-high-school-trigger-police-probe/.

[107] Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C. and Niessner, M. Face2face: Real-time face capture and reenactment of rgb videos. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, p. 2387–2395. Available at: https://openaccess.thecvf.com/content_cvpr_2016/html/Thies_Face2Face_Real-Time_Face_CVPR_2016_paper.html.

[108] Thing, V. L. Deepfake detection with deep learning: Convolutional neural networks versus transformers. In: IEEE. *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*. 2023, p. 246–253. Available at: https://ieeexplore.ieee.org/abstract/document/10225004.

[109] Tschandl, P.; Rinner, C.; Apalla, Z.; Argenziano, G.; Codella, N. et al. Human–computer collaboration for skin cancer recognition. *Nature medicine*. Nature Publishing Group US New York, 2020, vol. 26, no. 8, p. 1229–1234. Available at: https://www.nature.com/articles/s41591-020-0942-0.

[110] Tsigos, K.; Apostolidis, E.; Baxevanakis, S.; Papadopoulos, S. and Mezaris, V. Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection. In: *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*. 2024, p. 37–45. Available at: https://dl.acm.org/doi/abs/10.1145/3643491.3660292.

[111] Venkateswarulu, S. and Srinagesh, A. DeepExplain: Enhancing DeepFake Detection Through Transparent and Explainable AI model. *Informatica*, 2024, vol. 48, no. 8. Available at: https://www.informatica.si/index.php/informatica/article/view/5792.

[112] Vodrahalli, K.; Gerstenberg, T. and Zou, J. Y. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 2022, vol. 35, p. 4004–4016. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/hash/1968ea7d985aa377e3a610b05fc79be0-Abstract-Conference.html.

[113] Xu, G.; Hou, Y.; Liu, Z. and Loy, C. C. Mind the gap in distilling StyleGANs. In: Springer. *European Conference on Computer Vision*. 2022, p. 423–439. Available at: https://link.springer.com/chapter/10.1007/978-3-031-19827-4_25.

[114] Xu, Y.; Raja, K. and Pedersen, M. Supervised contrastive learning for generalizable and explainable deepfakes detection. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, p. 379–389. Available at: https://ieeexplore.ieee.org/abstract/document/9707568.

[115] Xu, Z.; Hong, Z.; Ding, C.; Zhu, Z.; Han, J. et al. Mobilefaceswap: A lightweight framework for video face swapping. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022, vol. 36, no. 3, p. 2973–2981. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/20203.

[116] Xue, Z.; Jiang, X.; Liu, Q. and Wei, Z. Global–local facial fusion based GAN generated fake face detection. *Sensors*. MDPI, 2023, vol. 23, no. 2, p. 616. Available at: https://www.mdpi.com/1424-8220/23/2/616.

[117] Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R. et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*. ACM New York, NY, USA, 2023, vol. 56, no. 4, p. 1–39. Available at: https://dl.acm.org/doi/full/10.1145/3626235.

[118] Yang, X. and Bo, H. High-Fidelity Face Swapping with Style Blending. *ArXiv preprint arXiv:2312.10843*, 2023. Available at: https://arxiv.org/abs/2312.10843.

[119] Yu, N.; Davis, L. S. and Fritz, M. Attributing fake images to gans: Learning and analyzing gan fingerprints. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, p. 7556–7566.

[120] Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F. and Hartley, R. Face super-resolution guided by facial component heatmaps. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, p. 217–233. Available at: https://openaccess.thecvf.com/content_ECCV_2018/html/Xin_Yu_Face_Super-resolution_Guided_ECCV_2018_paper.html.

[121] Zhang, N.; Luo, J. and Gao, W. Research on face detection technology based on MTCNN. In: IEEE. *2020 international conference on computer network, electronic and automation (ICCNEA)*. 2020, p. 154–158.

[122] Zhou, P.; Han, X.; Morariu, V. I. and Davis, L. S. Two-stream neural networks for tampered face detection. In: IEEE. *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. 2017, p. 1831–1839. Available at: https://ieeexplore.ieee.org/abstract/document/8014963.

[123] Zhu, Y.; Li, Q.; Wang, J.; Xu, C.-Z. and Sun, Z. One shot face swapping on megapixels. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, p. 4834–4844. Available at: https://openaccess.thecvf.com/content/CVPR2021/html/Zhu_One_Shot_Face_Swapping_on_Megapixels_CVPR_2021_paper.html.

Appendix A

Guidance

- **Information provided:**
  - **Deepfake detector result:** Provides the probability that the image is a deepfake. (E.g., 73% indicates a 73% chance the detector assessed it as a deepfake).
  - **Heatmap:** Shows the important parts of the face for the detector to decide.
    - **Red = Highest attention:** The redder the area on the map, the more the detector has caught its attention. Red areas indicate that the detector found distinctive features or characteristics in these areas of the face that it considered key to analyzing the image. These can be areas that are important for facial recognition in general, or specific details that the detector considers relevant.
    - **Blue = Lower attention:** Conversely, blue areas represent areas that the detector paid less attention to during analysis. It considers these areas to be less prominent or less important for its decision whether it is a real or fake face.
    - **Colors between red and blue:** Areas represent a gradual scale of attention and significance. The more the color moves towards red, the more weight the area has for the detector in the overall analysis of the image.
  - **Important Warning:** Deepfake detectors are not perfect and can be wrong. Do not consider their result as the absolute truth.
- **Task:** Your task is to decide whether the displayed face is **genuine** or a **deepfake**. You also need to indicate your level of **confidence** in this decision. When deciding, please consider the information provided by the detector (score and heatmap), but ultimately rely on your **own judgment**.
- **Conclusion:** Good luck!

Appendix B

Questionnaires

Initial Questionnaire:

- **Demographic Information:**
  - Age range
    (e.g., 18-24, 25-34, 35-44, 45-54, 55+, Prefer not to say)
  - Gender
    (e.g., Male, Female, Other, Prefer not to say)
  - Highest level of education attained
    (e.g., High school, Bachelor's degree, Master's degree, Doctoral degree, Prefer not to say)
  - Was it attained in technical sciences?
    (Yes, No, Prefer not to say)
- **Deepfake Experience:**
  - Have you ever created a deepfake image or video?
    (Yes, No, Prefer not to say)
  - Do you have any prior experience with deepfake detection tools or techniques?
    (Yes, No, Prefer not to say)
  - How confident are you in your ability to visually detect deepfake images?
    (Not at all, Slightly, Moderately, Very, Completely, Prefer not to say)
  - How familiar are you with the concept of „explainable AI“ or „interpretable AI“?
    (Not at all, Slightly, Moderately, Very, Completely, Prefer not to say)
  - In your opinion, how significant is the potential impact of deepfakes on society?
    (Not at all, Slightly, Moderately, Very, Completely, Prefer not to say)
  - How often do you encounter images or videos online that you suspect might be deepfakes?
    (Never, Not often, Sometimes, Often, Daily, Prefer not to say)

Post-Experiment Questionnaire (all questions are optional):

- How confident are you overall in the accuracy of your classifications?

  (Not at all, Slightly, Moderately, Very, Completely, Prefer not to say)

- (For Groups 2 and 3) How easy was it for you to understand the information presented?

  (Not at all, Slightly, Moderately, Very, Completely, Prefer not to say)

- (For Groups 2 and 3) How much did you rely on the AI detector's output when making your decisions?

  (Not at all, Slightly, Moderately, Very, Completely, Prefer not to say)

- (For Group 3) How helpful was the heatmap in making your decisions?

  (Not at all, Slightly, Moderately, Very, Completely, Prefer not to say)

- (For Group 3) If you answered anything other than „Not at all“ to the previous question, please describe how the heatmaps influenced your decision-making process. If you answered „Not at all“, please explain why you did not use the heatmaps in your decision-making.

  (Open-ended text box)

- Do you have any further comments or feedback about the experiment?

  (Open-ended text box)

Appendix C

Experiment Images

Examples of detector heatmaps overlaid on original images, along with the corresponding deepfake prediction scores.

Figure C.1: Heatmap example with 100% deepfake prediction score. Deepfake correctly classified as deepfake by detector.

0%

Figure C.2: Heatmap example with 0% deepfake prediction score. Genuine correctly classified as genuine by detector.

11%

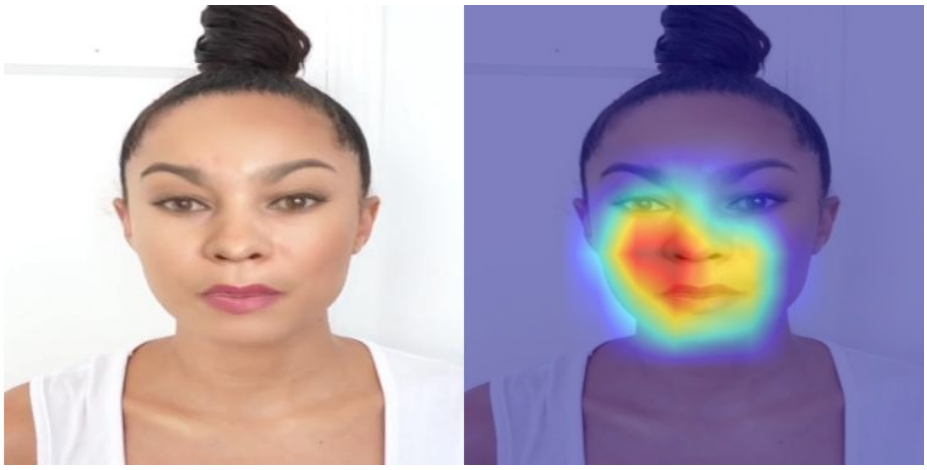


Figure C.3: Heatmap example with 11% deepfake prediction score. Deepfake incorrectly classified as genuine by detector.

76%

Figure C.4: Heatmap example with 76% deepfake prediction score. Genuine incorrectly classified as deepfake by detector.

Appendix D

Cloud storage content

The attached link to the cloud contains the thesis, including its source code LaTeX.

- **xbrnaf00.pdf** - this thesis,
- **xbrnaf00_source.zip** - source codes LaTeX to compile this thesis