Security Implications of Deepfakes in Face Authentication

Milan Šalko
Brno University of Technology
Brno, Czech Republic
isalko@fit.vut.cz

Anton Firc
Brno University of Technology
Brno, Czech Republic
ifirc@fit.vut.cz

Kamil Malinka
Brno University of Technology
Brno, Czech Republic
malinka@fit.vut.cz

ABSTRACT

Deepfakes are media generated by deep learning and are nearly indistinguishable from real content to humans. Deepfakes have seen a significant surge in popularity in recent years. There have been numerous papers discussing their effectiveness in deceiving people. What's equally, if not more concerning, is the potential vulnerability of facial and voice recognition systems to deepfakes. The misuse of deepfakes to spoof automated facial recognition systems can threaten various aspects of our lives, including financial security and access to secure locations. This issue remains largely unexplored. Thus, this paper investigates the technical feasibility of a spoofing attack on facial recognition. Firstly, we perform a threat analysis to understand what facial recognition use cases allow the execution of deepfake spoofing attacks. Based on this analysis, we define the attacker model for these attacks on facial recognition systems. Then, we demonstrate the ability of deepfakes to spoof two commercial facial recognition systems. Finally, we discuss possible means to prevent such spoofing attacks.

CCS CONCEPTS

• **Security and privacy** → **Access control**; **Spoofing attacks**; • **Social and professional topics** → **Spoofing attacks**.

KEYWORDS

deepfake, facial recognition, biometrics systems, machine learning, computer security

ACM Reference Format:
Milan Šalko, Anton Firc, and Kamil Malinka. 2024. Security Implications of Deepfakes in Face Authentication. In *The 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24), April 8–12, 2024, Avila, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3605098.3635953

1 INTRODUCTION

Deepfake is content created by artificial intelligence that is authentic in the eyes of humans but, in reality, depicts non-existing events or people. The term *deepfake* combines the words "deep learning" and "fake" and primarily refers to content created by deep neural networks, a branch of machine learning. The most common form of deepfakes involves creating and manipulating media associated with humans, such as face synthesis or face swapping. Their quality has advanced in recent years, making it often impossible for humans to distinguish deepfakes videos and photos from real ones [20, 23].

The increasing quality of fake copies is compounded by their increasing availability, which makes them accessible to a broader range of potential attackers. This confluence of factors reduces the technical expertise required, lowering the barriers to creating convincing deepfakes. As a result, even individuals with limited knowledge can create a relatively high-quality deepfake. The proliferation of user-friendly, open-source deepfake tools and pre-trained models supports this accessibility [10].

Nowadays, many articles [3, 12, 31] are explaining at length the dangerous consequences of misusing deepfake technology. In particular, significant attention has been paid to the malicious use of deepfakes to spread disinformation, political propaganda and other malicious purposes. In addition, the worrying potential of deepfakes as a means of damaging the reputation and credibility of individuals was highlighted with cases such as the creation of deepfake pornography [22]. Equally significant are the increasing fraudulent activities in which individuals pose as authority figures, such as corporate executives or family members, to gain illicit profits [7].

Within these critical discussions about the manifold implications of deepfakes misuse, one aspect that has received relatively limited attention so far is particularly important: the resilience of facial recognition systems to deepfake spoofing attacks. Recent studies [9, 27] have already confirmed that speaker recognition systems are vulnerable to deepfake spoofing attacks, but no special attention has been given to commercially deployed facial recognition systems to date. Facial recognition systems protect many areas where their exploitation could result in significant financial losses, such as an attack on smart banking solutions and personal or even national security through access to restricted or classified systems.

An incident involving a successful deepfake spoofing attack on facial recognition has already been reported in China in early 2021. Tax fraudsters used stolen facial images to create deepfake videos and a special phone with a hijacked camera to trick the tax invoicing system into accepting these pre-made deepfake identities to defraud $76.2 million [2].

The threat analysis of using deepfakes to spoof facial recognition is paramount. While it is known in the community that deepfakes can potentially deceive these systems, comprehending the specific scenarios and methods for implementing deepfake spoofing attacks is crucial. Not all use cases provide opportunities for deepfake spoofing attacks. For instance, scenarios involving supervision by a third party, like automatic border crossing with a border crossing officer, make it impractical to substitute a person's face with a deepfake at the gate and require a different approach, such as disrupting the passport creation process by providing a manipulated (morphed)

This work is licensed under a Creative Commons Attribution International 4.0 License.
SAC '24, April 8–12, 2024, Avila, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0243-3/24/04.
https://doi.org/10.1145/3605098.3635953

facial image. Similarly, systems that rely on a single image for verification can be tricked with a genuine photo of the same person, rendering deepfake usage unnecessarily complex.

Once we understand the nuances of these potential attacks, we can define the attacker model for deepfake spoofing. Ultimately, we aim to demonstrate the feasibility of these proposed attacks. While it is widely acknowledged that deepfakes can deceive facial recognition, providing proof within our established threat and attacker models adds a valuable layer of assurance.

In this work, we first perform a threat analysis to explore all facial recognition use cases that allow for a meaningful implementation of spoofing attacks. Secondly, based on the threat analysis results, we assess the ability of deepfakes to spoof two commercial facial recognition systems to validate if deepfake spoofing attacks present any threat. We examine the presentation of a single deepfake photo and a consecutive set of deepfake images extracted from a video. Finally, we propose a method to mitigate the threats posed by deepfakes to facial recognition systems. The main contributions of this paper might be summarized as follows:

- We conduct a threat analysis to identify scenarios in which deepfake spoofing attacks can be carried out within facial recognition applications and establish the attacker model.
- We assess the feasibility of deepfake spoofing attacks on two commercially deployed facial recognition systems and show they are vulnerable to such attacks.
- We demonstrate that even data of inferior quality is able to fool facial recognition systems to a certain extent.

Section 2 provides an overview of related work. Section 3 discusses facial recognition use cases and identifies use cases vulnerable to deepfake spoofing attacks. Experiment design is proposed in Section 4, and the results are provided in Section 5. Finally, Section 6 discusses the most interesting findings of this paper and proposes countermeasures, and Section 7 concludes the paper.

The work described in this document results from a previously completed master's thesis [33].

2 RELATED WORK

The topic of deepfakes is quite widespread in the field of research. This section describes several areas relevant to assessing the usability of deepfakes to spoof facial recognition systems and puts this work into the context of other relevant research.

Detection of facial deepfakes. As Mirsky et al. [20] state, much research deals with detecting facial deepfakes. Proposed methods often use different types of artefacts in the image, such as different flickers of illumination that do not fit the rest of the face that the creators are trying to detect [6, 11, 14]. The second common approach is to use deep learning to classify whether an image or video is a deepfake by learning its discriminative features [25, 26]. The last approach is tracking various physiological cues, such as blinking or a subtle change in skin colour due to a heartbeat [4, 5].

Deepfake datasets. Celeb-DF dataset [18] provides 639 high-quality videos of DeepFake celebrities. The advantage of this dataset is that it contains deepfake videos along with multiple real videos for each person, which, compared to other datasets. Another suitable dataset is KoDF [17]. It includes 62.8 days of records, which makes it the largest deepfake dataset, generated with the most widely known tools such as DeepFaceLab [19] or First Order Motion Model [28]. There are a large number of other datasets that are also known, for example, Face Forensic++ [24], DFDC (Deepfake Detection Challenge) [8] and FakeAVCeleb [15].

Feasibility of deepfake spoofing attacks. Previous works examined the usability of deepfakes to spoof facial recognition systems. However, to the best of our knowledge, none of these works examined deployed commercial systems. In addition, we define the use cases of such systems allowing meaningful implementation of deepfake spoofing attacks and evaluate whether using multiple consecutive video frames improves the resilience of facial recognition systems.

Tariq et al. [30] have attacked celebrity identification APIs from Microsoft, Amazon and Naver. These APIs have been shown to be deceived by deepfakes in up to 78% of cases. The experiments showed that some deepfake generation methods pose a more significant threat to recognition systems than others and that each system responds differently to attacks. The examined state-of-the-art models for celebrity identification cannot identify people who are not celebrities. It's important to note that the systems included in our testing are designed primarily for facial biometrics and personal identification of any individuals, not just celebrities.

Further work by Korshunov et al. [16] described that state-of-the-art facial recognition algorithms based on VGG and Facenet are vulnerable to Deepfake videos and cannot distinguish such videos from the original ones with up to 95.00% equal error rate. While these results are relevant, the work examines the resilience of the backbone networks instead of the whole systems.

3 USECASES OF FACIAL RECOGNITION

As mentioned earlier, to determine how well deepfakes can trick facial recognition systems, the initial step is to identify the particular scenarios where facial recognition technology is used and then identify potential situations where an attacker could execute a spoofing attack. This process sets the foundation for our threat modelling. Therefore, it is crucial to comprehend the various use cases of facial recognition and the likelihood of vulnerability to deepfake spoofing before commencing the test phase.

Facial recognition systems are an integral tool for identity verification in various areas, including airport security protocols and authentication processes for online services such as Internet banking. However, it is essential to note that not all use cases are vulnerable to being attacked using deepfake spoofing attacks. To provide a structured framework for understanding potential threats, we divide attacks on individual systems into three categories based on their vulnerability to deepfake spoofing attacks, as visually depicted in Figure 1.

Attacker model. An attacker is an individual with the capability to generate facial deepfakes with the intention of bypassing a system protected by facial recognition. This attacker possesses all necessary personal information (login data, birth number, or other type of identifier) about his target and details about the typical facial recognition process. Additionally, he has access to publicly available images and videos of the victim's face. Using this information, the attacker creates facial deepfakes of the victim and then attempts

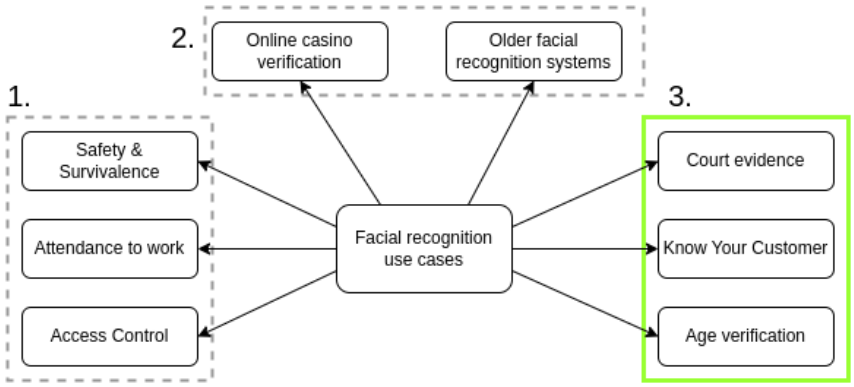

Figure 1: Examples of the use of facial recognition fall into three categories: 1. It is difficult to attack them with deepfakes. 2. It is not worth using deepfakes to attack. This category includes "Older facial recognition systems", which refers to all use cases of outdated and poorly designed systems for which more conventional attack methods are sufficient. 3. (green solid line) Appropriate use of deepfakes in an attack.

to gain access to the system secured by facial recognition in the most convincing manner possible. When he gets in, he will use the access granted to his advantage.

The attacker's objective is to carry out a deepfake spoofing attack. This attack involves impersonating a chosen victim and presenting this falsified identity to a facial recognition system. The ultimate goal is to authenticate the spoofed identity as the victim, thus obtaining unauthorized access to a system protected by facial recognition. The attacker has several methods to carry out such an attack. The most straightforward approach is to replay the deepfake from a device screen to the facial recognition capture device, as depicted in Figure 2. More sophisticated options include intercepting the data received from the capture device or gaining direct access to the API of the facial recognition system. It is important to note that these attacks differ from presentation attacks, as the attacker doesn't use actual samples from the victim; instead, they synthesize them. Additionally, there are distinctions from morphing attacks, where a deepfake is presented as the reference sample on an ID document. In a spoofing attack, the ID would contain an unaltered image, and the attacker would present the deepfake to the capture device at the border gate.

Use case categorization. The *first category* encompasses systems where deepfake spoofing is extremely difficult or unfeasible. This includes use cases within security forces or services, such as border crossings, representing a specialised access control form. These entities are equipped to promptly detect and respond to suspicious attempts to spoof the deployed facial recognition technology. For instance, in the context of automated border crossing gates at airports, where facial recognition is commonly used to verify travellers' identities, deepfakes pose a significant challenge and are highly unlikely to succeed. This same difficulty level applies to other scenarios with monitoring and control mechanisms to identify abnormal or deceptive behaviour during facial recognition.

It's worth noting that deepfakes do not threaten safety and surveillance systems as well. This is because deepfake attacks require the attacker to employ an electronic device that conceals their

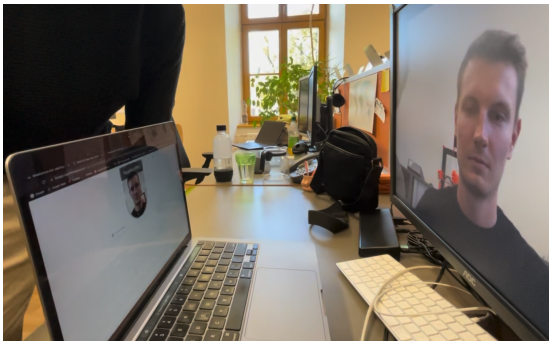

Figure 2: Example of replaying deepfake from a device screen to the facial recognition capture device to execute a deepfake spoofing attack.

face while projecting the deepfake identity in the monitored area. This approach is both impractical and overly complex; as such, a disguise would be promptly detected. Attendance systems at workplaces also fall into this category for similar reasons. In most cases, security personnel oversee all employees entering the building, making it nearly impossible to present deepfakes to the attendance system without being detected. Additionally, 3D facial recognition systems are part of this category. They are designed to inherently resist 2D spoofing attacks, including deepfakes, making them less susceptible to manipulation by standard 2D deepfake techniques. Moreover, we are not aware of any 3D deepfake creation techniques.

The *second category* includes use cases where deepfake spoofing attacks are pointless. Such use cases include outdated facial recognition systems vulnerable to basic presentation attacks, where a static photo of the victim is sufficient for authentication. Similarly, poorly designed systems that require only a single photo or document for authentication might be spoofed using more straightforward and more conventional attack vectors, such as presenting an image of the victim without any modification, making deepfakes unnecessary and redundant. These use cases also include online casinos

that only require the user to upload a photo ID and one face photo from a mobile phone.

The *third category* includes use cases well suited for deepfake spoofing attacks. Deepfakes are better suited to exploit systems requiring a more robust video-based authentication process, making them a targeted choice for a subset of biometric security scenarios. Ideal use cases for deepfake attacks are those where real-time human oversight is lacking and input data for verification requires a more complex format based on multiple images or videos. Notable examples that meet these criteria include Know Your Customer (KYC) systems, particularly in online banking. In such scenarios, a successful deepfake attack could give the attacker unauthorised access to the victim's bank account.

The absence of immediate human intervention in these online authentication processes creates a favourable environment for an attacker to carefully prepare and execute an attack without arousing suspicion.

A good example of such an environment is age verification. An attacker can use freely available tools to age the face in a photo. He then creates a video from this photo using one of the few-shot deepfake techniques. Using this video, he then verifies his age, allowing him to access a social network or purchase age-restricted goods.

Another relevant example is in court evidence, where facial recognition is used to match suspects' identities with evidence material. Deepfakes, thus, might be used to fabricate false accusations. Submitting a deepfake video would incriminate innocent victims in order to cause harm or evade justice. In such cases, the attacker benefits from the ability to create deepfake content carefully in advance and then submit it for analysis.

In summary, using deepfake technology to spoof facial recognition systems becomes a strategic choice for attackers if certain conditions are met. In particular, deepfake attacks are most relevant when the target system requires a higher level of authentication that requires the submission of video-based authentication instead of accepting a simple photo. It is also ideal if no other authority supervises the authentication process and the attacker has enough time to prepare for such an attack or repeat it several times.

4 EXPERIMENTAL DESIGN

In the previous sections, we briefly outlined the essence of testing the resilience of facial recognition to deepfake spoofing attacks and defined vulnerable use cases and the attacker model. Based on this knowledge, this section describes in detail the experiment design for assessing the resilience of facial recognition to deepfake spoofing attacks. As mentioned in Section 3, we focus on use cases that require a video clip to be submitted and then extract one or more frames from this video for verification. The experiments are divided into two parts: the first examines a scenario where just a single frame is extracted and used for verification, and the second examines a scenario where multiple consecutive frames are extracted and used for verification.

For both phases of the experiment, individual images and image sequences will be selected from the Celeb-DF dataset. Only suitable images that meet the quality requirements are selected. These requirements are described later in the experiment design. These data are fed into the selected Megamatcher and IFace biometrics, which represent state-of-the-art commercial facial recognition systems. For both experiment phases, three types of scores are collected: impostor, genuine, and deepfake.

From these scores, distribution graphs and individual metrics such as false non-match rate (FNMR) and false match rate (FMR) are calculated. The statistical similarity is computed between the different types of scores to determine whether they depend on each other.

There seems to be an apparent mismatch between the previously defined use cases and how the selected facial recognition systems work. However, we can simplify all the described use cases to a simple vector comparison problem due to the essence of how facial recognition applications work. The biometric system (facial recognition) performs a basic operation in which it extracts a vector containing the key points of the face from the input face image. This vector is then compared with the pattern vector (user profile) stored in a database or provided as a reference identity. The result of this vector comparison is a similarity score, telling us if tested persons are of the same identity. All of the defined use cases are only built on these foundations. Thus, if we omit the use case-specific information unrelated to the facial recognition technology, all of the identified use cases can be simplified into this vector comparison problem.

The first part thus operates with the simplest option: image-to-profile comparison. Only a single facial image (frame) is extracted from the input video in this setting. This image is then compared to the user profile stored in the database or reference image containing the user identity. We execute multiple attempts to record:

- **Genuine score** - comparison scores will be collected for each identity by comparing mated samples (images of the same person – identity).
- **Impostor score** - comparison scores will be collected for each identity by comparing non-mated samples (images of different persons – identities).
- **Deepfake score** - comparison scores will be collected for each identity by comparing deepfake samples with genuine samples of the same identity (deepfake of identity X with deepfake impersonating identity X).

We utilize this score to generate distribution plots and compute error rates. Specifically, we focus on the FNMR and FMR in this testing scenario. These rates are determined by applying various matching score thresholds to the data. The point at which the FMR and FNMR curves intersect is the equal error rate (EER). The EER value informs us that if we set the acceptance threshold t, the False Accept and False Reject rates will be equal, resulting in an error rate of x. A lower EER indicates better system performance.

Given our consideration of deepfakes as impostors, meaning we aim to prevent their access to the system, we may use them to generate EER plots instead of impostor scores. These new plots are instrumental in gauging the impact of presenting deepfakes on the system's performance. Finally, we use statistical tests to examine if the differences between the genuine, impostor and deepfake scores are statistically significant.

The second phase of our approach operates in a more advanced setting. It involves comparing multiple frames extracted from the

video with the user profile stored in the database or the reference image containing the user's identity. The scores for each frame are then averaged to yield the final score. This approach might offer greater resilience against deepfake spoofing attacks, as it can identify inconsistencies across multiple frames. The recorded scores will be assessed in the same manner as in the first part.

4.1 Selected biometric systems

For the experiments, two commercial facial recognition systems are used. The first is the IFace SDK 3.0 [29]. Its features include real-time identification and authentication (1:1 matching), multi-face tracking, and person analysis, including age and gender profiling. The technology is based on deep neural networks and provides verification capabilities from still images and video footage in all standard formats. Two facial images are fed into the system, between which the similarity is calculated. The application can work in two modes:

- *Fast mode* - some partially obscured faces or faces with sunglasses may be overlooked. Also, faces printed on ID cards may not be recognized. However, the speed performance of face recognition is much better than that of other modes.
- *Accurate mode* - Partially obscured faces with blurred profiles or faces with sunglasses are recognized. The processing speed of this face detection is lower than that of the other modes.

The second system is Megamatcher [21]. It is designed for developers of large-scale AFIS (Automatic Fingerprint Identification System) and multi-biometric systems, available as a software development kit that enables the development of large-scale products for the identification of one or more biometric features such as fingerprint, iris, face, voice or palm print. The SDK is available for Microsoft Windows, Linux, macOS, iOS and Android platforms. This technology ensures high reliability and speed of biometric identification even when using large databases. The creators of Megamatcher provide a web interface as a demo application and a trial version of the SDK that can be run in a terminal. For our experiments, we used the terminal version. In contrast to IFace, Memagamtcher allows for the creation of user profiles and verification against these profiles instead of comparison between two facial images.

4.2 Dataset Preparation

As previously mentioned, our experiments use the Celeb-DF [18] dataset. The dataset contains 58 identities. Its main advantage is that it always provides ten videos of a real person and 30 videos of deepfakes. For this reason, it is the most suitable data source. However, none of the tested systems supports video embedding but image embedding; we had to extract single frames from each video and verify their suitability for further testing. One of the identified disadvantages is that the videos in the dataset are collected from various talk shows, so often, the person is not looking directly into the camera. To overcome this problem, we tried to make the extracted frames in the dataset correspond as much as possible to the ICAO standard [1]. To achieve this, we check multiple parameters of each frame and discard the unsuitable ones.

First, we detect the face itself in the image. For this, we use MTCNN implementation [32], which can compute the key points of the face in addition to the region itself. We use these extracted points to determine whether the person is looking directly into the camera. Once we obtain the face region, we only select images where the person has a neutral facial expression. For this, we use the pre-trained model [1] which determines the facial expression of a person. Only images where the person has a neutral facial expression are used.

For the first part of the experiments, we selected the best single frame regarding compliance with the ICAO standards. For the second part, a sequence of frames is required. We thus select ICAO-compliant frames that are at least one second apart. This condition helps to eliminate selecting identical frames next to each other.

5 EXPERIMENT RESULTS

As mentioned, the experiments are divided into two parts. The first part examines image-to-profile comparison; the second examines image sequence-to-profile comparison. The results show how resilient tested facial recognition systems are to deepfake spoofing attacks and whether image sequence-to-profile provides more resilience.

5.1 First part: image-to-profile comparison

After performing all the comparisons for both face recognition systems, we could compute all three required scores: genuine, deepfake and impostor score. From these scores, distribution function plots were then created. The distribution plots and EER graphs for each tested system are shown in Figure 3 (IFace accurate mode), Figure 4 (IFace fast mode), and Figure 5 (Megamatcher).

The plots clearly illustrate a disturbing phenomenon - the overlap between deepfake and genuine scores. This overlap essentially confirms that some deepfake attempts have the potential to score high enough to fool systems into accepting them as genuine. This has significant implications for the security and authenticity of digital media.

In addition, it is also worth noting that only a minority of deepfake attempts have been categorised as imposters. Deepfake scores reside in between impostors and genuine ones. This may be especially problematic if we determine the system threshold only using the EER values obtained from impostors and genuine attempts. In such cases, the threshold for the IFace system in both settings would be slightly below 20% similarity. Such a threshold setting would allow most genuine attempts to succeed, and impostor attempts to be rejected, as documented by the EER value very close to zero. However, if deepfakes are presented to the system with such a setting, the majority of deepfakes would be accepted, as evidenced by the distribution plots, which creates a big security issue.

Finally, we perform the statistical analysis using an independent Student's t-test. Using a significance level $\alpha = 0.05$, the test proves no significant similarity between all score types. While this means that deepfake scores are significantly different from the genuine ones, they are also significantly different from impostor ones. This

[1] https://github.com/Azure/MachineLearningNotebooks/blob/master/how-to-use-azureml/deployment/onnx/onnx-inference-facial-expression-recognition-deploy.ipynb

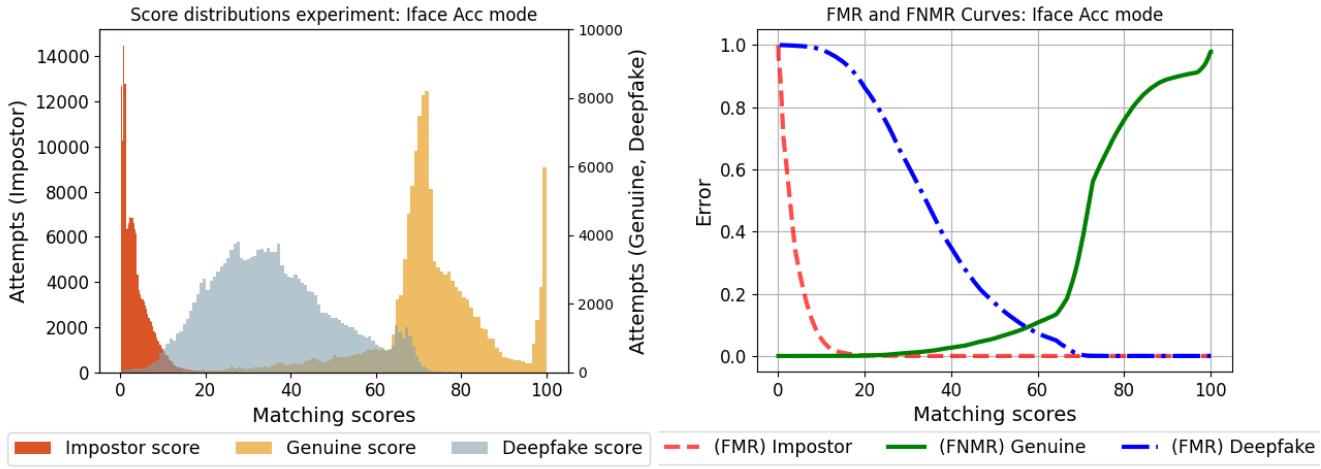

Figure 3: Matching scores distribution graph on the left and right FMR / FNMR graphs for the IFace accurate mode for image-to-profile comparisons.

supports our claim that deepfake scores reside between genuine and impostor ones, which may cause significant security issues.

In summary, this experiment revealed that some comparisons achieved results that reached a sufficient threshold to accept deepfakes as valid input. Thus, if we required that the system's security be kept at a level that would prevent an attacker from getting in using a deepfake, we would have to raise the threshold to a level at which even a fraction of genuine inputs would be rejected. This shows the potential vulnerability of these systems to deepfakes attacks.

5.2 Second part: image sequence-to-profile comparison

In the second part of the experiment, we selected sequences of five images from the suitable images selected during dataset preparation. These were spaced one second apart. The final score is an average of the five obtained scores.

Similar to the preceding section, Figure 6 illustrates that the EER between the impostor and genuine scores is nearly zero. Surprisingly, there's no notable change in the EER between deepfake and genuine scores. This suggests that the system struggled to identify deepfakes even with multiple snapshots. Consequently, the facial recognition system did not exhibit the expected increase in resilience to deepfakes.

To highlight this behaviour, we compared the EER plots for image-to-profile with those for image sequence-to-profile and found no discernible difference in the EER values. This further supports our theory that the chosen facial recognition systems struggle to detect deepfakes as invalid inputs on their own effectively.

The statistical analysis, conducted using an independent Student's t-test with a significance level of $\alpha = 0.05$, yielded the same results as in the first part. There is no statistically significant similarity between the score types. Thus, while the system may perform well with a low threshold setting on real data, it remains highly vulnerable to deepfake threats.

In summary, this experiment demonstrates that utilizing image sequences for facial recognition does not enhance resilience against deepfakes. The observed increases in EER between genuine-impostor and genuine-deepfake scores mirrored those observed in the first part.

6 DISCUSSION

Our research results suggest potential vulnerabilities of facial recognition systems to deepfake technologies. This section discusses various aspects of these results and suggests countermeasures to mitigate the identified threats.

6.1 Quality of deepfakes in dataset

The experiments used the older Celeb-DF dataset, which no longer reflects the current capabilities of deepfakes. This difference can be seen in Figure 7, which compares the deepfake from Celeb-DF and the current GHOST [13] model. To address this challenge, we could have used the KoDF dataset, which contains superior-quality deepfakes. However, this dataset was collected in a controlled environment where all subjects spoke into a camera. Conversely, Celeb-DF introduces much more variability into the data, better reflecting real-world conditions. Breaching the systems using older methods for creating deepfakes indicates the low requirements for implementing this type of attack.

It is also important to note that this dataset was not created as a dataset to test the robustness of biometric facial recognition systems to deepfake materials. This fact is particularly evident in the quality of the videos, which were obtained from freely available sources on the Internet. These videos do not fully meet the requirements for identity verification based on ICAO standards [1].

Regarding the dataset's quality, it is essential to consider capturing the similarity between the victim and the face actor in the context of face swapping. This is particularly important because it is very likely that an attacker seeking to gain unauthorized access to the system will try to ensure that the deepfake material is of the

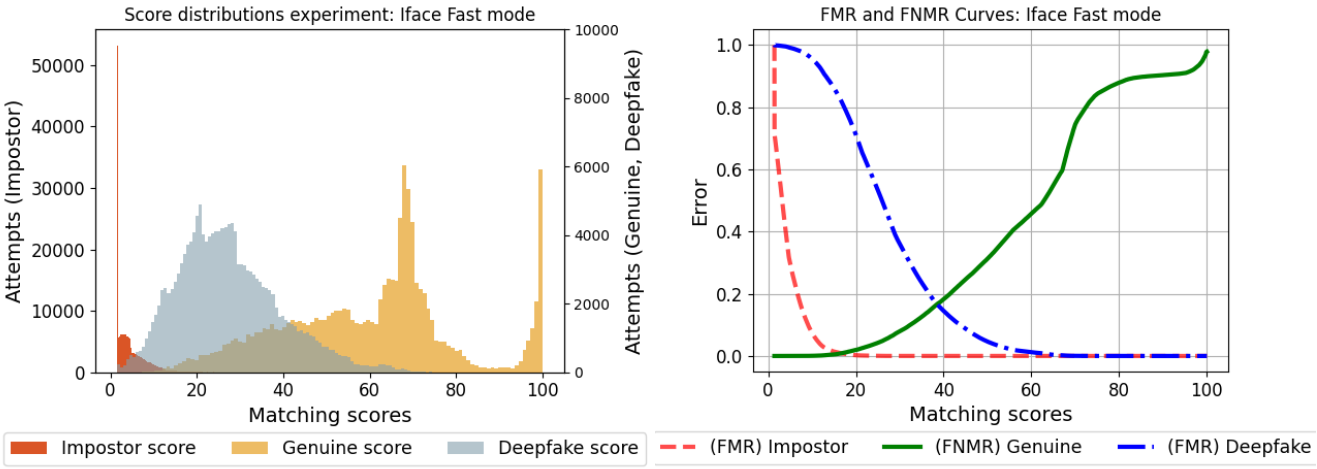

Figure 4: Matching scores distribution graph on the left and right FMR / FNMR graphs for the IFace fast mode for image-to-profile comparisons.

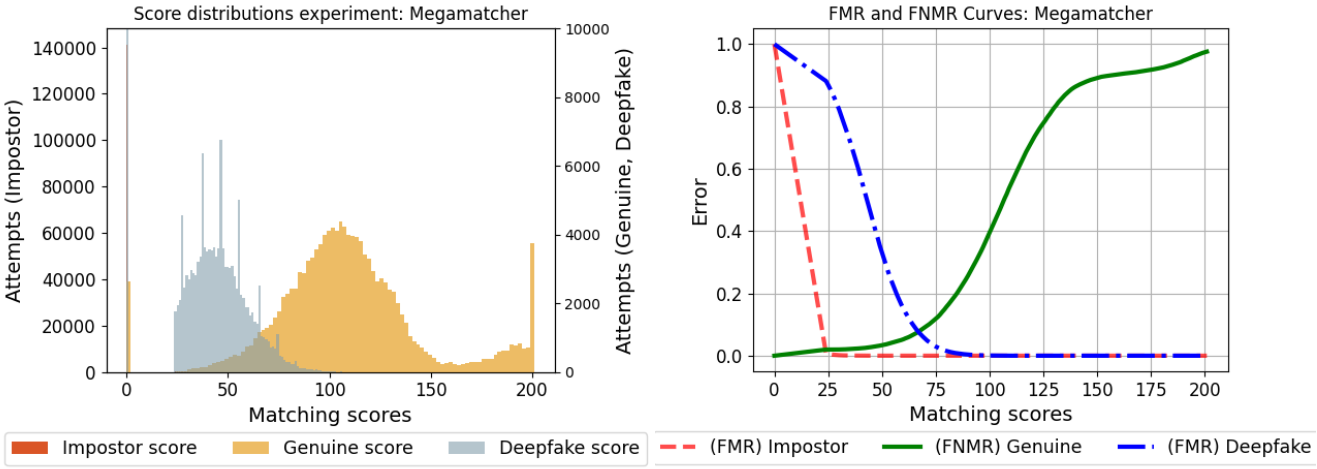

Figure 5: Matching scores distribution graph on the left and right FMR / FNMR graphs for the Megamatcher for image-to-profile comparisons. Megamatcher automatically assigns a zero score to all impostor attempts.

highest quality he or she is able to achieve. However, the lower quality of the deepfakes used does not limit the impact of this work. We demonstrate that even data of inferior quality is enough to spoof state-of-the-art commercial facial recognition systems. Additionally, the success rate of deepfake spoofing attacks only increases with better-quality deepfakes.

Therefore, a dataset that reflects current deepfake creation techniques and also satisfies the requirements on the quality of the resulting image for biometric systems should be developed in the future. This dataset should focus on generating deepfakes that closely mimic the victim's appearance, including parameters like skin tone, hair colour, and facial features.

6.2 Countermeasures

As we show, deepfakes present a significant threat to facial recognition systems. It is thus essential to implement countermeasures that mitigate the posed threats. During our experiments, we observed an essential shortcoming of state-of-the-art deepfake creation tools that might be exploited as a liveness detection method to diminish the usage of deepfake to spoof facial recognition reliably. As shown in Figure 8, current deepfake creation tools struggle when an object passes in front of the face of the deepfake actor. In the worst cases, the whole *deepfake mask* disappears for a portion of time, revealing the original actors' identity.

This imperfection might be currently used as a very effective means to spot deepfake videos. The liveness test would ensure that the verified subject would, for example, wave his hand in front of

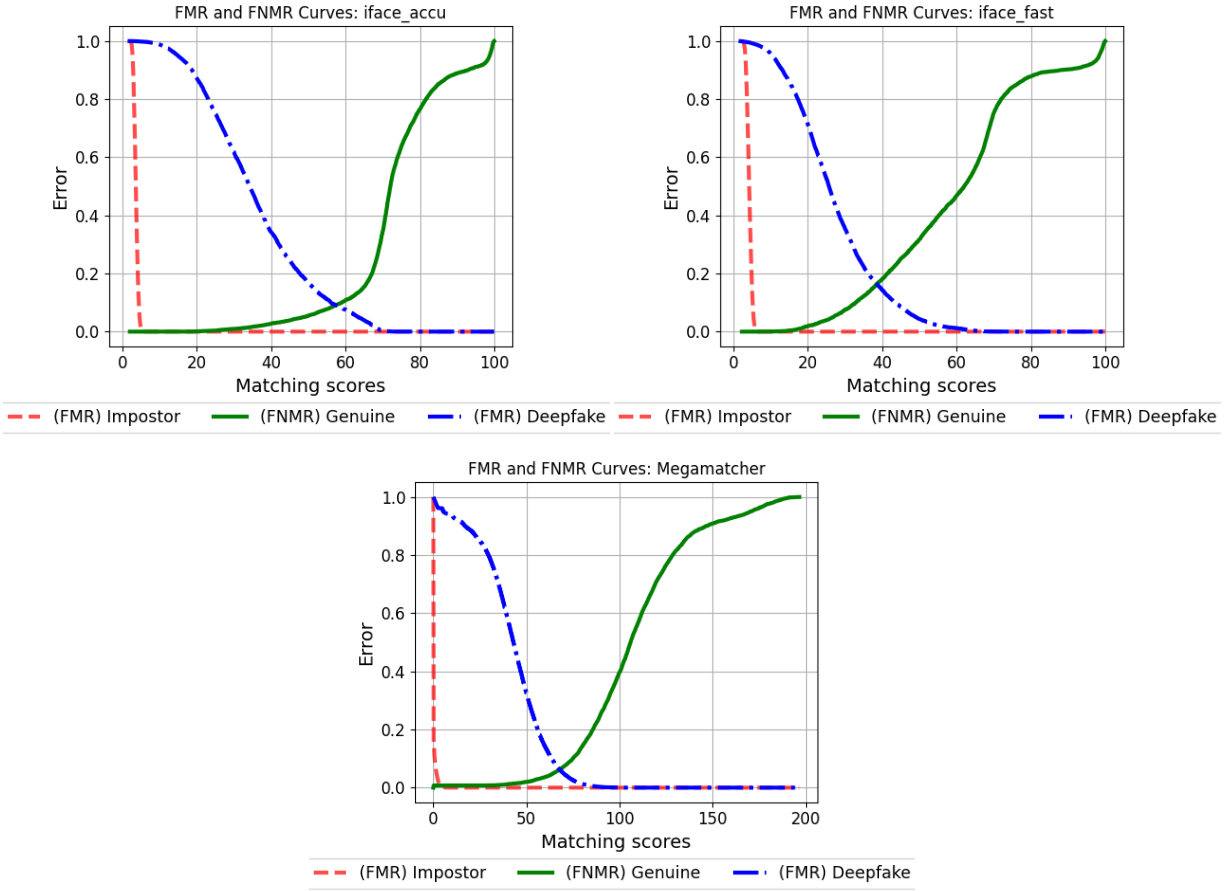

Figure 6: FMR / FNMR charts from top iFace (fast mode), iFace (accurate mode) and Megamatcher for image sequence-to-profile comparisons.

Figure 7: Comparison of the quality of deepfakes in the Celeb-DF dataset with the quality of current tools (GHOST [13]).

Figure 8: Example of distortion when an object passes in front of the face of a deepfake actor.

the face and simultaneously look for any artefacts or inconsistencies caused by this movement.

We understand that it is only a matter of time before this problem is resolved, but this approach should provide rapid and reliable mitigation for the next few years to alleviate the bulk of the threats posed. In the meantime, there is enough space to develop more generalizable mitigation solutions and deploy them before the proposed solution becomes ineffective.

7 CONCLUSION

The targeted use of deepfake techniques to spoof facial recognition is most relevant in contexts that require video-based authentication, where a single photographic image of the person under examination

can no longer be relied upon. And those where the attacker has ample time and preparation to carry out such an attack.

Our findings show that the tested systems failed to identify a certain fraction of deepfake inputs as impostor inputs. This highlights the problem that deepfakes can pose for these systems. Considering the scope of use of these systems, this is a significant problem. Moreover, in our work, we worked with a relatively old Celeb-DF dataset, but we could still spoof the tested facial recognition systems. This also means the problem is likely even more pronounced when modern techniques are used. These results highlight the need to update and improve deepfake detection systems.

Future work should, therefore, focus on testing other facial biometric systems using modern tools for creating deepfakes.

This potential vulnerability may be even more pronounced when biometric facial recognition systems are used. Only in the cases we have mentioned could there be significant financial losses or the wrong person could be apprehended. We believe this paper succeeds in drawing attention to the growing problem of using deepfake technologies to spoof automatic facial recognition systems.

ACKNOWLEDGMENTS

This work was supported by the Brno University of Technology internal project FIT-S-23-8151. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

[1] BioPass ID. [n. d.]. Learn the requirements of the ICAO standard (ISO 19794-5). https://www.biopassid.com/post/norma-icao

[2] Masha Borak. 2021. Tax scammers hack government-run facial recognition system. https://www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax

[3] Johnny Botha and Heloise Pieterse. 2020. Fake news and deepfakes: A dangerous threat for 21st century information security. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited*. 57.

[4] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[5] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). https://doi.org/10.1109/tpami.2020.3009287

[6] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. 2020. On the Detection of Digital Face Manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5780–5789. https://doi.org/10.1109/CVPR42600.2020.00582

[7] Audrey de Rancourt-Raymond and Nadia Smaili. 2022. The unethical use of deepfakes. *Journal of Financial Crime* 30, 4 (May 2022), 1066–1077. https://doi.org/10.1108/JFC-04-2022-0090

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. 2020. The DeepFake Detection Challenge Dataset. *CoRR* abs/2006.07397 (2020). arXiv:2006.07397 https://arxiv.org/abs/2006.07397

[9] Anton Firc and Kamil Malinka. 2022. The Dawn of a Text-Dependent Society: Deepfakes as a Threat to Speech Verification Systems. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* (Virtual Event) *(SAC '22)*. Association for Computing Machinery, New York, NY, USA, 1646–1655. https://doi.org/10.1145/3477314.3507013

[10] Anton Firc, Kamil Malinka, and Petr Hanáček. 2023. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon* 9, 4 (2023), e15090. https://doi.org/10.1016/j.heliyon.2023.e15090

[11] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. 2021. Fighting Deepfakes by Detecting GAN DCT Anomalies. *Journal of Imaging* 7, 8 (Jul 2021), 128. https://doi.org/10.3390/jimaging7080128

[12] Chandell Gosse and Jacquelyn Burkell. 2020. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication* 37, 5 (2020), 497–511. https://doi.org/10.1080/15295036.2020.1832697 arXiv:https://doi.org/10.1080/15295036.2020.1832697

[13] Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, and Denis Dimitrov. 2022. Ghost—a new face swap approach for image and video domains. *IEEE Access* 10 (2022), 83452–83462. https://doi.org/10.1109/access.2022.3196668

[14] Jun Jiang, Bo Wang, Bing Li, and Weiming Hu. 2021. Practical Face Swapping Detection Based on Identity Spatial Constraints. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*. 1–8. https://doi.org/10.1109/IJCB52358.2021.9484396

[15] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. 2021. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=TAXFsg6ZaOl

[16] Pavel Korshunov and Sebastien Marcel. 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. arXiv:1812.08685 [cs.CV]

[17] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. 2021. KoDF: A Large-Scale Korean DeepFake Detection Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10744–10753.

[18] Yuezun Li. 2020. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*.

[19] Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. 2023. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition* 141 (2023), 109628. https://doi.org/10.1016/j.patcog.2023.109628

[20] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (jan 2021), 41 pages. https://doi.org/10.1145/3425780

[21] Neurotechnology. 2023. Megamatcher SDK. https://www.neurotechnology.com/megamatcher.html

[22] Person and Shane Raymond. 2021. Deepfake anyone? Ai Synthetic Media Tech enters perilous phase. https://www.reuters.com/technology/deepfake-anyone-ai-synthetic-media-tech-enters-perilous-phase-2021-12-13/

[23] Daniel Prudký, Anton Firc, and Kamil Malinka. 2023. Assessing the Human Ability to Recognize Synthetic Speech in Ordinary Conversation. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 1–5. https://doi.org/10.1109/BIOSIG58226.2023.10346006

[24] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *International Conference on Computer Vision (ICCV)*.

[25] Ulrich Scherhag, Christian Rathgeb, and Christoph Busch. 2018. Towards Detection of Morphed Face Images in Electronic Travel Documents. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. 187–192. https://doi.org/10.1109/DAS.2018.11

[26] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. 2017. Detection of Face Morphing Attacks by Deep Learning. In *Digital Forensics and Watermarking*. Springer International Publishing, Cham, 107–120.

[27] John Seymour and Azeem Aqil. 2018. Your Voice is My Passport. https://www.blackhat.com/us-18/briefings/schedule/#your-voice-is-my-passport-11395

[28] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf

[29] Innovatrics s.r.o. 2021. Face functions. https://developers.innovatrics.com/digital-onboarding/docs/functionalities/face/

[30] Shahroz Tariq, Sowon Jeon, and Simon S. Woo. 2022. Am I a Real or Fake Celebrity? Evaluating Face Recognition and Verification APIs under Deepfake Impersonation Attack. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 512–523. https://doi.org/10.1145/3485447.3512212

[31] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[32] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[33] Milan Šalko. 2023. Security Implications of Deepfakes in Face Authentication. Available at https://www.vut.cz/en/students/final-thesis/detail/141060.