Realistic Fingerprint Presentation Attacks Based on an Adversarial Approach

Roberto Casula, *Member, IEEE*, Giulia Orrù, *Member, IEEE*, Stefano Marrone, Umberto Gagliardini, Gian Luca Marcialis, *Senior Member, IEEE*, and Carlo Sansone, *Senior Member, IEEE*

***Abstract*— Modern Fingerprint Presentation Attack Detection (FPAD) modules have been particularly successful in avoiding attacks exploiting artificial fingerprint replicas against Automated Fingerprint Identification Systems (AFISs). As for several other domains, Machine and Deep Learning strongly contributed to this success, with all recent state-of-the-art detectors leveraging learning-based approaches. An insidious flip side is represented by adversarial attacks, namely, procedures intended to mislead a target detector. Indeed, despite this type of attack has been considered unrealistic, as it presupposes access to the communication channel between the sensor and the detector, in a recent work, we have highlighted the possibility of transferring a fingerprint adversarial attack from the digital domain to the physical one. In this work, we take a step further by introducing a new procedure designed to make the physical adversarial presentation attack i) more robust to the physical crafting of the PAI by exploiting explainability techniques, ii) easier to adapt to different fingerprint scanners and adversarial algorithms, and iii) usable in a black-box scenario. To quantify the impact of these novel adversarial presentation attacks family, designed to be robust to the physical crafting process, we assess the performance of both state-of-the-art PAD modules alone and integrated AFISs. Results highlight the approach's feasibility, opening a new series of threats in the context of fingerprint PAD.**

***Index Terms*— Fingerprint, adversarial, presentation attack, AFIS.**

I. Introduction

RECENT years have seen a substantial increase in Automated Fingerprint Identification Systems (AFISs) accuracy. This biometry can fairly be considered the most academically and industrially mature and is one of the most appropriate options for any application requiring a high level of security [1]. This success is mainly due to its universality, persistence, and uniqueness. However, AFISs are vulnerable to presentation attacks (PA), the submission of artificial replicas to the biometric sensor, with the aim of impersonating an authorized user [2]. This type of attack is prevalent in real-world biometrics applications, as attackers do not need access to the recognition system's internal modules. Fingerprint Presentation Attack Detection (PAD) [3] methods are used to counteract this possibility by classifying the images acquired with the sensor in bona fide, *i.e.* belonging to a real finger, or PA, *i.e.* obtained through an artificial replica. PADs usually base their decision on measurement and analysis of anatomical, physiological or texture-based characteristics extracted from the fingerprint images [4]. The advent of deep learning has made PAD systems even more accurate than human expert analysis. However, the use of deep-learning increases the vulnerability to adversarial attacks, which are already present for generic machine learning-based systems. An adversarial attack against PADs is the possibility of modifying the classification outcome by digitally perturbing the input image [5]. Fortunately, this type of attack is realized in the digital domain and requires the attacker to be able to access the communication channel between the sensor and the classifier. Access to the communication channel is difficult to obtain. For this reason, adversarial attacks are not as widespread as presentation attacks that target the biometric sensor. However, it has recently been shown that adversarial attacks can be presented to the biometric sensor when used as the basis of the construction of the Presentation Attack Instrument (PAI) for both attacking facial [6] and fingerprint recognition systems [7]. The physical realization of the PAI, starting from the adversarial images and their presentation to the sensor, allows attacking the system directly from the "exposed" component. In the following, we refer to this type of attack as "adversarial presentation attack" (ADV-PA), that is, an adversarial attack carried out in the physical domain.

However, the physical realization of the adversarial PA can alter information about the user (in particular, the position and characteristics of the minutiae), making the attack on the AFIS useless. In this work, we proposed a new perturbation generation technique called *"Focus Attention"* to obtain a digital perturbation robust to the printing process that does not alter the fingerprint user-specific characteristics. In particular, we adopted an iterative perturbation method based on image processing techniques to highlight the information of the fingerprint, namely ridges and valleys, and modify just the pixels within it since such changes result after printing and PAI re-acquisition. The obtained perturbation is used to provide novel PAIs. Since both digital adversarial attacks and adversarial presentation attacks are claimed to require

Manuscript received 14 June 2023; revised 2 October 2023 and 18 October 2023; accepted 21 October 2023. Date of publication 25 October 2023; date of current version 23 November 2023. This work was supported in part by the Piano Nazionale Ripresa Resilienza (PNRR) Ministero dell'Università e della Ricerca (MUR) Project under Grant PE0000013-FAIR. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhen Lei. *(Corresponding author: Giulia Orrù.)*

Roberto Casula, Giulia Orrù, and Gian Luca Marcialis are with the Department of Electrical and Electronic Engineering, University of Cagliari, 09123 Cagliari, Italy (e-mail: roberto.casula@unica.it; giulia.orru@unica.it; marcialis@unica.it).

Stefano Marrone, Umberto Gagliardini, and Carlo Sansone are with the Department of Electrical Engineering and Information Technologies, University of Naples Federico II, 80125 Naples, Italy (e-mail: stefano.marrone@unina.it; u.gagliardini@studenti.unina.it; carlosan@unina.it).

Digital Object Identifier 10.1109/TIFS.2023.3327663

© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

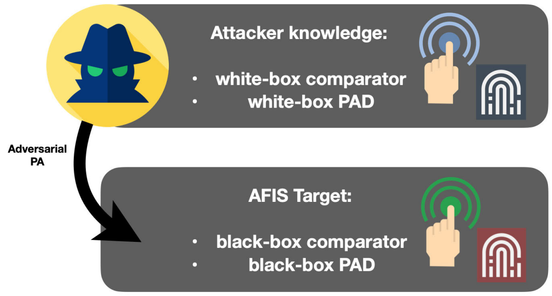


Fig. 1. Transfer of a presentation adversarial attack designed on a known AFIS to a completely unknown AFIS.

knowledge about the comparator and PAD modules to compute a proper perturbation and the manufacturers of AFISs for today's personal or public safety devices hide the implementation details, we assessed the actual threat level of the proposed ADV-PA testing it on "black-box" AFISs. We show that if the "Focus Attention" PAI is submitted to the sensor of a black-box PAD, there is a high possibility that it will be misclassified as a bona fide fingerprint *despite the attacker having zero knowledge about the target system*.

Figure 1 simulates a realistic attack context where the attacker knows neither the comparator nor the PAD of the targeted AFIS. We stress that it is not required that the same sensors be adopted during the attack design (on white-box systems) and implementation (on black-box systems) [3]. The rest of this paper is organized as follows: Section II makes an overview of the current literature in order to explain the different ways of attacking an AFIS and adequately collocate the proposed analysis; Section III describes the purpose of the experimentation and the proposed method; The experimental methodology and results are presented in Section IV; Conclusions are drawn in Section V.

II. Related Work

A. Vulnerability of Integrated Authentication Systems

A biometric system can be attacked in all its components (Fig. 2). The possible attacks are to the communication channels among the modules (points 2, 4 and 6 in the figure), such as replay attacks and hill-climbing attacks, to the specific modules via malware infection (points 3, 5, 7, 9 in the figure) or to the template database via template theft, substitution, or deletion (point 8) [8]. Digital adversarial attacks compromise the channel at point 2. However, as the sensor is the most exposed part of AFISs, PAs are the most concrete risk (point 1).

For this reason, PADs [9] are in charge of managing the threats due to the presentation of artificial fingerprint replicas, also known as presentation attacks (PAs). Over two decades, scientific and applied research proposed new techniques for the design of PADs [10], [11]. In recent years, PADs based on deep learning have become the most common due to their very high accuracy and the availability of data and computational resources [12]. However, the use of these methods has accentuated pre-existing vulnerabilities, such as the possibility of deceptively modifying the PAD decision with the perturbation of a few pixels. These attacks, called adversarial attacks, are described in detail in the following Section.

B. Adversarial Attacks

In [13] the authors showed that deep-learning models can be easily fooled into making a wrong prediction by imperceptibly perturbing target samples with noise. These noise injections are known as adversarial perturbations or adversarial attacks. This vulnerability has led researchers to develop many attack techniques that can be classified into white-box and black-box attacks. Their difference lies in the knowledge of the attackers. White-box attacks assume that the adversary has complete knowledge of the targeted model and none in black-box attacks. Over the years, numerous adversarial attack algorithms have been proposed, most of them white-box, such as limited-memory Broyden–Fletcher–Goldfarb–Shanno's (L-BFGS) [13], the fast gradient sign method (FGSM) [14], the basic iterative method (BIM)/projected gradient descent (PGD) [15], distributionally adversarial attack, Carlini and Wagner (C&W) attacks [16], Jacobian-based Saliency Map Attack (JSMA) [17], and DeepFool [18]. The Fast Gradient Sign Method (FGSM) is the first method that uses the network gradient to generate an additive adversarial attack [14] multiplying a user-defined ϵ, serving as the adversarial perturbation's ∞-norm bound, by the sign of the prediction gradient. Based on the functioning of the FGSM, other methods have been developed including the FGSM Iterative Method, the DeepFool and the Momentum Iterative Method. More specifically, DeepFool [18] computes a more optimal adversarial example, approximating the smallest possible perturbation to reverse the classification. It is based on an efficient iterative approach that exploits the network gradient of a localized version of the loss. Other methods of generating gradient-based attacks include C&W, JSMA and PGD. In particular, PGD [15] is an attack based on the multiple executions of FGSM until an incorrect classification is obtained, which allows for generating a perturbation that maximizes the loss of a model on a particular input while keeping the size of the perturbation less than a specified amount called epsilon. A couple of years later, a new version of the attack was released under the name of Auto Projected Gradient Descent (APGD) [19]. As the name suggests, it improves the standard PGD by auto-optimising the attack strength across iterations before restarting and re-running the attack from the best-found example. Another noteworthy adversarial attack is the One-Pixel Attack [20], based on Differential Evolution, an evolutionary optimisation method that changes one or a few pixels to mislead the classification of a network. It is important to note that the One-Pixel attack does not need access to the target CNN's white-box and may be conducted without any previous network knowledge.

The reported list is not intended to be exhaustive but just to report those that, over the years, contributed the most to the domain's advance, highlighting that several different ways can be pursued to perform an adversarial attack.

1) Adversarial Attacks to Biometric Authentication Systems in the Digital Domain: Recently, it has been shown that it is possible to adapt adversarial perturbation to CNN-based biometric recognition. An AFIS was attacked first by [5]. The authors proposed to breach an AFIS equipped with a

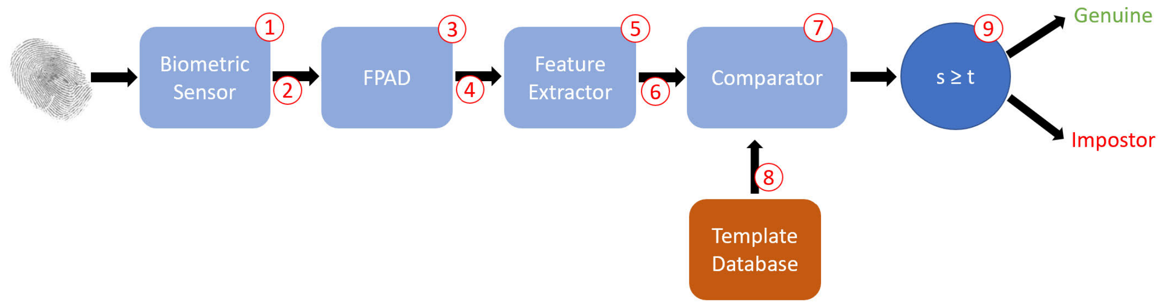


Fig. 2. Illustrative schema for the possible vulnerable points of a biometric authentication system (fingerprint-based in the example). The scanner (1) acquires the biometric data (2) and sends it to the presentation attack detection module (3). The response (4) is used by the feature extraction module (5) to decide whether to elaborate the input biometrics or reject it. In the former case, extracted features (6) are compared by the comparator (7) against the set of features extracted for authorised users (8). Based on the stages' outputs, the comparator module (9) calculates a similarity score s between them, comparing it to a decision threshold t in excess of which the sample is deemed a mated trial. Given this schema, the possible attacks are to the communication channels among the modules (points 2, 4 and 6), such as replay attacks and hill-climbing attacks, to the specific modules via malware infection (points 3, 5, 7, 9) or to the template database via template theft, substitution, or deletion (point 8). However, as the sensor is the most exposed part of AFISs, PAs are the most concrete risk (point 1).

PAD module starting from an artificial replica of the victim's fingerprint. Three adversarial methods are used: FGSM, DeepFool and One-Pixel Attack. The methods, albeit with different results, have proved effective in piercing integrated AFISs based on different acquisition sensors. Nonetheless, it is worth noting that this first proof-of-concept might not be able to mislead an expert as they tend to introduce distortions not common in fingerprint images.

On the same line, [21] analysed the transferability of some adversarial perturbation attacks against face recognition systems. The results showed that even "naive" attacks (i.e., performed without adapting the attack to face recognition) can be successfully transferred under specific settings. Further analyses are reported in [22] and [23].

However, all these attacks are white-box. More recently, in [24], the authors took a first step toward a completely black-box attack, albeit in the digital domain, by examining whether transferring a perturbation between different CNN PADs is possible. Although these attacks have proved to be successful, they are also purely theoretical: they presuppose direct access to the comparator or to the PAD module, which, if available, can be used directly with an original sample "stolen" from the victim (a face photo or a fingerprint capture). In other words, they need to attack, at the same time, the AFIS and the communication channel.

2) Adversarial Attacks to Biometric Authentication Systems in the Physical Domain: Transferring digital adversarial attacks into the physical domain, that is, fabricating a PAI from adversarial images, means converting a theoretical risk into a real risk. In the following, we refer to this as *adversarial presentation attack* or *ADV-PA*. The first transfer of an adversarial attack into the physical domain was against a face authentication system equipped with a deep learning-based PAD in 2020 [6], demonstrating that it is possible to cheat an integrated face recognition system in a white-box context by submitting to the sensor biometric traits modified by adversarial perturbations.

In the fingerprints field, the first case was presented by [7], in which a DeepFool white-box attack against a CNN-based PAD showed the feasibility of an ADV-PA and highlighted how some variables of the acquisition process affect the final result. In particular, to carry out the physical attack, a layer of liquid latex was dripped over the prints of the perturbed fingerprints and once dry, it was presented to the acquisition sensor.

This paper represents the follow-up of that early publication [7], with the following significant additions:

- We introduced a novel perturbation generation technique designed to optimize the attack, keeping the fingerprint's user-specific information unchanged and adapting it as best as possible to the printing process. In particular, three different versions of two adversarial methods, namely APGD and DeepFool, were presented. The novel perturbation generation technique is based on a focus attention mechanism capable of aligning the perturbation and the ROI of the white-box PAD. This mechanism, which is detailed in the next Section, allows the perturbation to be concentrated on the image areas engaged in the PAI realization.
- The fingerprints chosen to carry out the perturbation come from a different sensor than the one used to acquire the corresponding physics. This choice was made to compare the detectors presented at the LivDet 2019 and LivDet 2021 competitions with the same sensor. In this manner, a cross-sensor analysis is performed.
- In addition to the effectiveness of the attack on standalone PADs, we also evaluated the effectiveness on an integrated system to evaluate whether the new generative method allows keeping the information used by the comparator to authenticate the individual.
- The experimental study has been supplemented by entirely black-box tests to assess the threat of the attack in a realistic context.

III. Adversarial Presentation Attacks: The Focus Attention Method

Fingerprint adversarial PAs are difficult to carry out because, if not properly designed, the printing process can destroy the perturbations introduced into the image. Furthermore, the fingerprint image has unique characteristics that should not be altered because they are critical in the authentication process.

The authentication of the fingerprint is based, in particular, on the analysis of local ridge features known as minutiae. Moreover, highlighting the noise between the ridges of the fingerprint or that present in the image's background must be avoided. Our previous works [7] on adversarial PA were limited to printing a digital adversarial attack without taking the above critical issues into consideration.

This paper proposes a new approach for generating a customized perturbation to create realistic fingerprint adversarial PAs. Three distinct versions have been proposed to determine which procedure is more robust to the printing process. We tested different white-box and black-box AFISs under different experimental protocols. The goal is to assess these attacks' risks and acquire knowledge to prevent them.

Section II pointed out that realising an ADV-PA is not trivial because of the high number of involved factors that make the result reliable, reproducible and effective. In the case of a *fingerprint adversarial presentation attack*, the main difficulties to be taken into account are:

- dealing with the nature of fingerprint images, where all the information is located in a reduced portion of the whole image and is characterised by extremely domain-specific patterns, unique for each subject (i.e., images very different from other classification problems where the target classes are objects, animals, road signs, etc.);
- taking into account how the PAD is operating: (i) are only sub-portions of the whole image processed (*e.g.*, by focusing on minutiae-centred patches)? (ii) are pre-processing techniques to reduce noise introduced during the fingerprint acquisition process?;
- designing a suitable crafting strategy able to physically cast the PAI without destroying the injected adversarial alterations as a result of the crafting operation itself;
- to be of a practical appeal, the attack should operate in a black-box scenario, namely without any prior knowledge about the targeted AFIS.

To cope with these aspects, we introduce an iterative multi-stage fingerprint adversarial presentation attack (ADV-PA) equipped with a mechanism we called **Focus Attention** (FA). This aims to focus the adversarial perturbations on the image portion where the distinguishing features are found. The ADV-PA process is based on the concept of "shadow" presentation attack detector [24] (i.e., the known white-box detector used to craft the adversarial samples), and leveraging the experience we matured in printing latex-based PAIs [7], [25]. The idea aims to generalize the concept of adversarial perturbation of fingerprint images since, whichever the PAD functioning is, it is reasonable to hypothesize that the salient features are located along ridges, which must also be reproducible when printing the adversarial image. The proposed attack stages (Fig. 3) are the followings:

1) An attacker replicates the targeted fingerprint and presents it to the acquisition scanner. The attack is unsuccessful because the PAD module correctly classifies the PA and blocks authentication;
2) The fingerprint identified in the previous step undergoes the fingerprint adversarial perturbation stage, an iterative process to create adversarial images that mislead the targeted PAD, leveraging the shadow detector. This stage is detailed in Section III-A;
3) If successful, the digital adversarial fingerprint is physically cast, obtaining a physical replica or PAI (PAI). This stage is detailed in Section III-B;
4) The PAI is acquired by using the fingerprint scanner.

It is worth noting that all the process is designed to be black-box. Indeed, Stage 2 uses the known white-box PAD approach only to determine a suited adversarial perturbation without any prior knowledge about either the fingerprint scanner or the PAD used in Stage 4. Similarly, Stage 3 prints the fingerprint having the same size as the original image without focusing on the particular fingerprint scanner used afterwards. The next sections detail the core procedure (Stages 2-3), including the rationale behind our choices and a final practical description of the physical printing and casting process.

A. Digital Attack

As described in Section II-B, adversarial fingerprints are harder to realise than adversarial natural images due to their peculiarities in shape and patterns. This is even harder if the realised adversarial fingerprint has to be robust to the physical casting process while preserving the subject authentication characteristics. Unfortunately, the two aspects are in contrast, as the former would benefit from stronger/wider injected perturbations, which are more likely to destroy the subject's minutiae. Moreover, to make the process as general as possible and potentially easier to adapt to future developments, since PAD is an arms-race problem, we also want the adversarial fingerprint attack to be based on standard adversarial perturbation algorithms [16], [17], [18]. To deal with these needs, we designed an iterative process consisting of the following stages:

1) As adversarial algorithms need a target classifier to be attacked, the attacker trains a CNN-based PAD to the aim. In the following, we refer to this with the term *white-box PAD* to specify that the attacker knows every implementation detail of this classifier. It is worth highlighting that this PAD is not the actual target of the attack (as we are in a total black-box scenario) but a different PAD trained by the attacker with the aim of supporting the development of the adversarial attack. For this reason, we referred to it as *shadow classifier* [24] when we first introduced it, as the PAD based on it is crafted by the attacker in a fully white-box term setting as an adaptation of the substitute technique to the presentation attack detection application. The white-box PAD is used to determine the adversarial perturbation to be injected into the targeted fingerprint image. As this serves as the attacker's starting point, it is of crucial importance to maximise its detection rate as much as possible. In fact, a weak detector is easily misled, causing the adversarial fingerprints not to be effective against a better PAD, as explained by [24];
2) The targeted fingerprint image is resized to match the white-box PAD input size. For example, if the white-box

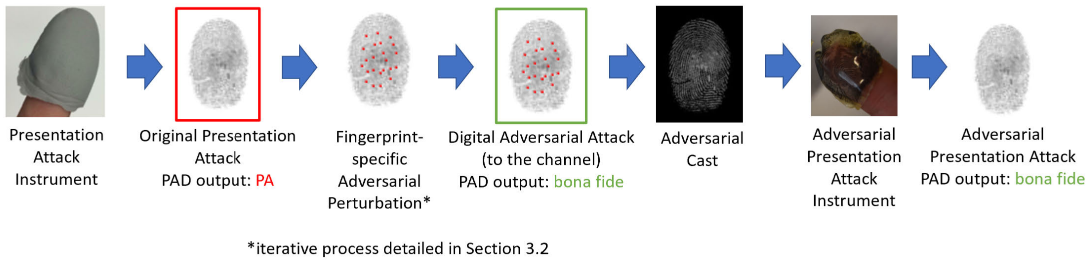


Fig. 3. Stages implementing a fingerprint ADV-PA, starting from a PA sample classified as PA by the target PAD.

PAD requires an image of $N \times M$ as input, the targeted fingerprint image will be resized with this size;

3) The resized sample undergoes the iterative adversarial perturbation process until it is recognised as bona fide. In particular, whatever adversarial perturbation algorithm is adopted:
   a) The RGB adversarial perturbation is determined for the whole image by using the white-box PAD as a target. Please note that this step is iterative itself;
   b) The RGB perturbation is masked to remove all the perturbed pixels belonging to the background, obtaining the perturbation ROI. To produce the mask (Fig.4), opening and closing with a 5px-diameter circular structural element followed by Otsu's thresholding were used, as proposed in [5];
   c) The masked RGB perturbation is converted to greylevels, where R, G and B are the colour channels. It is worth noting that this procedure results in a single-channel perturbation. Thus, if the considered white-box CNN-based PAD expects a 3-channel input, the so-obtained greylevel perturbation can be replicated over all the channels;
   d) The perturbation is injected into the targeted fingerprint, and the obtained adversarial perturbed fingerprint is tested against the considered PAD. Points 1-2-3 are repeated until one of the following stopping conditions is met: i) a fixed maximum number of iterations is reached ($ITER\_MAX$); ii) the adversarial attack is successful (*i.e.*, the predicted class is "bona fide" in our case). The latter condition is very important as, usually, an attack is considered successful as soon as the class changes. However, as in our case we need a more robust perturbation, we change the stopping condition by adding a value c^*, acting as "bona fide" class probability threshold. The value of c^* has been set to 0.8 based on [24], while $ITER\_MAX$ is automatically selected by the algorithm, with a maximum possible value of 200 (very high, used just as a cap) and an early stopping strategy based on a patience of 50 (set as a cap to reduce the computational effort in the case of semi-plateau regions);

4) If the attack in the previous point is successful, the perturbed fingerprint image is resized back to the original fingerprint size (the one before point 1).

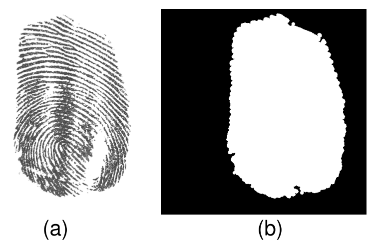


Fig. 4. The original fingerprint (a) and the corresponding binary mask (b), obtained by using morphological operators and Otsu's threshold.

The pseudo-code for the whole procedure is reported in algorithms 1-2. In particular, algorithm 2 describes the white-box classifier training procedure. Although in this work we used DeepFool [18] and APGD [19] as adversarial techniques, it is worth noting that *the whole method is independent of the perturbation algorithm adopted.*

Algorithm 1 Pseudo-Code for the Black-Box Adversarial Perturbation of Fingerprints. Italic Words Represent Variables, While Bold Ones Functions and Procedures. The Adversarial Method Can Be Any of Those Described in This Paper or Others Not Analyzed. The White-Box PAD (*white_pad*) Is Trained on a Dataset Acquired With the Same or a Different Sensor Than the Target AFIS. The PAD Takes as Input Images of *white_pad.InputSize* Size. *targ_fing* Is the Target Fingerprint From Which the Adversarial PA Is Generated. *targ_fing.OriginalSize* Is the Target Fingerprint's Original Size. *succeed* and *failed* Are Two Keyword Used to Simplify the Concept of "Digital PAD Attack Successful" and "Digital PAD Attack Failed" Respectively. *adv_fing* Is the Obtained Adversarial Perturbed Fingerprint Image.

```
procedure GENERATE_ADVERSARIAL_FINGERPRINTS(targ_fing, white_pad)
    white_pad ← Train_WhiteBox_Classifier
    resize targ_fing to white_pad.InputSize
    adv_fing, succeed ← Fingerprint_Adversal_Perturbation(targ_fing, white_pad)
    if succeed then
        resize adv_fing to targ_fing.OriginalSize
        return adv_fing
    end if
return failed
end procedure
```

B. From a Digital Attack to a Printable Cast: The Focus Attention Mechanism

The proposed attack returns excellent results in the digital domain, but the crafting process destroys the introduced

Algorithm 2 Pseudo-Code of the **Fingerprint_Adversarial_Perturbation** Used Within the Algorithm 1. In Particular, It Describes the Specific Process of Generating an Adversarial Perturbation on a Target Fingerprint (*targ_fing*). Italic Words Represent Variables, While Bold Ones Functions and Procedures. The Adversarial Method (**Adversarial_Perturbation**) Can Be Any, Including Those Described in This Paper and Others Not Analyzed. The Adversarial Perturbation (*rgb_adv*) Is Applied Only to the Pixels Belonging to the ROI (*fp_mask*) Containing the Fingerprint to Avoid Enhancing the Noise in the Background. The Perturbation Process Is Iterated Until the Probability of Being Bona Fide of the Perturbed Image Is Greater Than a User-Specified Confidence Value *c* or After *ITER_MAX* Iterations

```
procedure FINGERPRINT_ADVERSARIAL_PERTURBATION(targ_fing, white_pad)
    fp_mask ← Extract_Fingerprint_ROI(targ_fing)
    adv_pert ← zeros(white_pad.InputSize)
    iter ← 0
    repeat
        iter++
        rgb_adv ← Adversarial_Perturbation(targ_fing, white_pad)
        masked_rgb_adv ← rgb_adv * fp_mask
        masked_rgb_adv_gray ← RGB2GRAY(masked_rgb_adv)
        adv_pert ← adv_pert + masked_rgb_adv_gray
        adv_fing ← targ_fing + adv_pert
        live_prob ← white_pad.Classify(adv_fing)
    until iter > ITER_MAX OR live_prob > c*
    succeed ← 0
    if live_prob > c* then succeed ← True
    end if
return adv_fing, succeed
end procedure
```

perturbation, making the attack unusable. Indeed, some of the so-obtained adversarial details are not replicable during the crafting processes described in Section III-C. However, the generation of the adversarial attack is an iterative process that includes the classification of each iteration outcome through the white-box classifier. We can take advantage of these repetitions to verify (and then enforce) that the perturbation procedure always operates within the same PAD's Region of Interest (ROI, i.e., the only portion of the image where the fingerprint is actually located). To the aim, we evaluate the behaviour of the PAD adopted in the design process (the white-box PAD) using a XAI technique [26] known as Occlusion [27]. This is a perturbation-based approach that allows evaluating the contribution of each pixel of the input image for a two-class classification model, producing an attribution map reproducing the original input to which a different colour into a green-red range is attributed according to the final classification.

Some examples of attribution maps for some fingerprint images are reported in Fig. 5, where pixels in green contribute positively to the activation of the target output and lead the network to decide on the predicted class (whatever is, "PA" or "bona fide"), while red ones lead the network towards the opposite class. The attribution maps show that the background pixels often influence the classifier prediction, especially for images classified as "PA". To evaluate the influence of background pixels on the classification of fingerprint images, in Table I we reported the percentage of significant pixels outside the ROI obtained through the Occlusion method on LivDet2015 test images. This analysis confirms that the adopted PAD tends towards the PA class when it finds information outside the ROI. We claim this is due to the "filth" present in the background, especially among the valleys introduced by the acquisition of an artificial replica. This clearly may lead to a misalignment between the PAD's ROI and the perturbation's ROI: if the perturbation affects the background, there is no way to provide a successful PAI. Thus, a very effective digital attack would be only because of the false correlation induced by the network and the image's label. Therefore, we need to nullify the role of background and, in general, of out-of-fingerprint pixels.

TABLE I

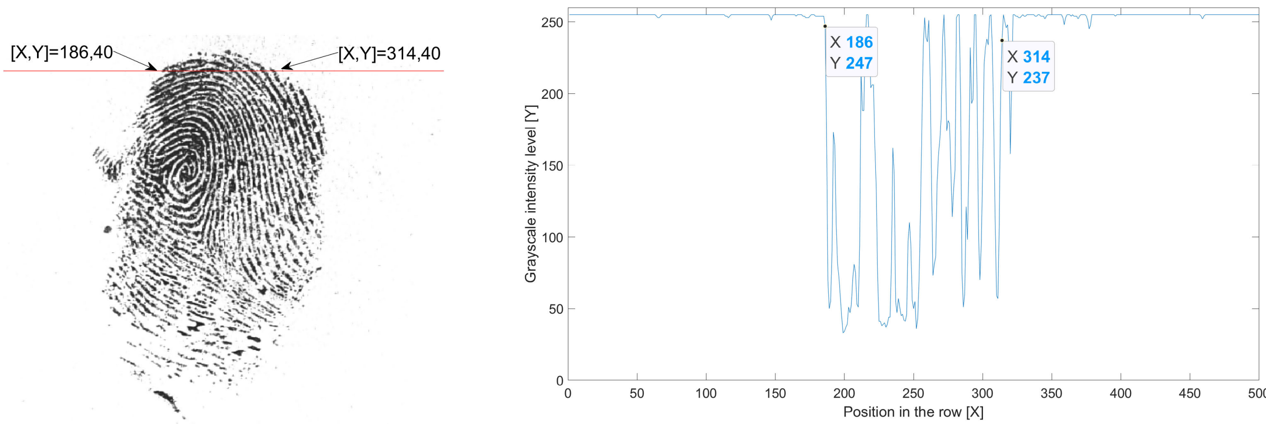
MEAN AND STANDARD DEVIATION OF THE NUMBER OF SIGNIFICANT PIXELS OUTSIDE THE ROI ON THE LIVDET2015 TEST SET USING THE OCCLUSION EXPLAINABILITY METHOD ON THE WHITE-BOX PAD [28]

	Percentage of significant pixels outside the ROI [*mean* $\pm$ *std dev*]
BF samples classified as BF	65.53 ± 17.22
BF samples classified as PA	75.75 ± 20.61
PA samples classified as PA	90.93 ± 11.74
PA samples classified as BF	50.03 ± 7.19

A possible solution is to create the equivalent of a constant signal for those pixels, such that, during the perturbation generation, the gradient is strongly reduced. Consequently, pixels belonging to the fingerprint ROI are the only ones leading the classification, and the perturbation is, in turn, focused on that region only. In contrast, unifying the pixel values in the background areas allows to prevent them from affecting the classifier's decision. In particular, if we represent the fingerprint image as a set of piecewise signals (Fig. 6), i.e. one signal for each row or for each column, the dirt in the background and in the valleys constitute spikes of low intensity. The elimination of these small spikes that randomly contribute to the gradient can be done by thresholding the intensity level of the image pixels, as these are characterized by low-intensity levels. In this work, we achieve this through a binarization process. It is important to underline that with binarization, not only are the values above a certain threshold brought to 255 but the values below the threshold are brought to 0. This results in a better definition of the ridges and valleys. Figure 7 shows the attribution maps of the same images after binarization. It is evident how binarization leads the classifier to analyze the portion of the image containing information, as the dirt in the background is eliminated. In other words, the binarization effect is to align the ROI of the white-box classifier with the ROI of the perturbation, what we call **Focus Attention** mechanism. This is a crucial pre-processing stage before the crafting stage, as the focus attention mechanism allows concentrating the perturbed pixels within the fingerprint regions. These regions will then be crafted as explained in the following and may retain the altered pixels in the foil.

(a) Bona fide sample classified as bona fide (b) Bona fide sample classified as bona fide (c) PA sample classified as PA (d) PA sample classified as PA (e) PA sample classified as PA

Fig. 5. Explainability through occlusions method on the VGG-CNN white-box network for bona fide and PA images. The attribution maps highlight in green pixels that contribute positively to the activation of the target output and in red pixels that suppress it.



(a) (b)

Fig. 6. Analysis of the 40th row (choice as an example) of a fingerprint image (a), marked with a red line. Pixels belonging to the real fingerprint image's texture start at pixel 168 and end at pixel 314. Dirt in the background and between the ridges produces small spikes in the intensity level signal (b).

Otherwise, they would be dispersed in the background and lost during crafting.

Starting from this modelling, three different versions of the attack have been proposed, each intended to produce images more robust to the crafting process by different image processing techniques. It is worth noting that the binarization phase is always done in all the variants, as it helps to eliminate noise that would be enhanced in the crafting phase and is crucial to shift the attention of the classifier to the ROI of the fingerprint. The following versions have been implemented for both DeepFool and APGD methods:

- *Focus Attention* adversarial presentation attack (FA): The binarization process is applied before and after each adversarial iteration;
- *Uniform Focus Attention* adversarial presentation attack (UFA): In general, perturbations are primarily applied in certain areas of the image depending on the gradient value. If the crafting process corrupts the portion of the image where the attack is concentrated, the perturbation is lost, and the adversarial attack is not effective. For this reason, at each iteration, the perturbation is applied to a different area of the ROI of the fingerprint in order to obtain uniformly distributed perturbations. Also in this case, the binarization process is applied before and after each adversarial iteration;
- *Robust Focus Attention* adversarial presentation attack (RFA): The binarization process often makes the ridges and minutiae too thin. The crafting process could corrupt fine lines, changing the fingerprint structure and making the comparison process based on the analysis of the minutes impossible. Therefore, a binarization step followed by a dilation step is applied before and after each adversarial iteration.

Therefore, six versions of PAIs for ADV-PAs have been developed: the Focus Attention, the Uniform Focus Attention, and the Robust Focus Attention based on DeepFool perturbations are respectively referred to with the terms FA-DF, UFA-DF and RFA-DF, whilst the Focus Attention, the Uniform Focus Attention, and the Robust Focus Attention based on APGD are respectively referred to with the terms FA-APGD, UFA-APGD and RFA-APGD.

C. Adversalial PAI Crafting

For the realization of the PA, starting from the digital adversarial image, the moulds of the adversarial fingerprints are created. The perturbed image is inverted to obtain the negative and printed on a translucent sheet using a standard laser printer, keeping the PA's original size. Then, a layer of latex is placed over the prints, ensuring that there is no air swelling and that the resulting layer has a suitable thickness for accurate removal and subsequent acquisition through the sensor (Fig. 9). As shown in Fig. 10, which contains the acquisition of two replicas of the same fingerprint

(a) Bona fide sample classified as bona fide (b) bona fidesample classified as bona fide (c) PA sample classified as PA (d) PA sample classified as bona fide (e) PA sample classified as bona fide

Fig. 7. Explainability through occlusions method on the VGG-CNN white-box network of bona fide and PA binarized fingerprint images. The attribution maps highlight in green pixels that contribute positively to the activation of the target output and in red pixels that suppress it.

(a) (b) (c)

Fig. 8. Image processing in the process of creating the digital adversarial PA in order to make it robust for printing: starting from the original image (a), binarization (b) is applied after each iteration of each version of the attack. In version 4 of the attack, dilatation (c) is applied after binarization.

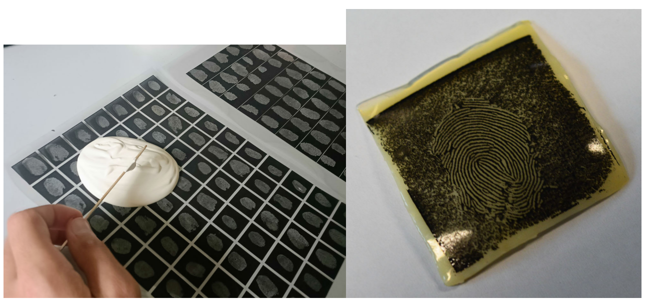


Fig. 9. Adversarial PA realization: the material is poured on the sheet with one or more prints of perturbed fingerprints. Once dried (24 to 72H), each replica is detached, cut out and presented to the biometric sensor.

Fig. 10. PAs of the same fingerprint obtained from two castings with different material thicknesses.

with different thicknesses of material, the skill and experience with the material by the operator is essential to obtain a good PAI. A realistic PAI, maintaining the adversarial perturbations after the print, is obtained by preliminary tests; the smallest printable detail and the best thickness of the latex on the sheet are found, and the final PAI is provided.

IV. Experiments

A. Datasets

In the following experimental analysis, we adopted LivDet 2015 [29], LivDet 2019 [30] and LivDet 2021 [12] Green Bit data sets. These datasets consist of bona fide and PA fingerprint images. The three datasets were acquired with two different sensors: LivDet 2015 with Green Bit DactyScan 26 and LivDet 2019 and LivDet 2021 with Green Bit DactyScan 84c. The composition of the datasets is shown in Table II. In particular:

- The LivDet 2015 training set was used to train the white-box PAD;
- The black-box PADs are pre-trained on LivDet 2019 and LivDet 2021 training sets;
- The adversarial PAIs were created from the latex samples of LivDet 2015 GreenBit. The adversarial PAs are cross-sensor because the original samples were acquired with Green Bit DactyScan 26 and, after the application of the perturbation and the re-print, were acquired with Green Bit DactyScan 84c;
- The LivDet 2019 and LivDet 2021 test sets were used as the baseline for the error rates of the black-box systems.

B. Experimental Protocol

To verify the potential of adversarial PAs, as a first analysis, we started with a white-box AFIS, consisting of a Bozorth3 comparator[1] in series with the VGG-CNN PAD [28] (Fig. 11) and a set of PA images correctly classified as PA (as an already successful PA does not require an adversarial perturbation). As shown in the reported diagram (Fig. 11), the white-box AFIS foresees a first check by the PAD module [31]; only if this check is passed, i.e. the fingerprint image is classified as "bona fide", then the comparison is evaluated via state-of-the-art NIST Bozorth. The white-box protocol represents the design phase of the attack from the attacker's point of view. For this reason, the VGG-CNN PAD, winner of the LivDet 2015 competition, has been selected as the white-box PAD: it is simply implementable by fine-tuning the well-known and open-source VGG neural network pre-trained on natural images. In this way, we acted as an actual attacker who uses an open-source, white-box PAD to obtain perturbed images,

[1] https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis

TABLE II

COMPOSITION OF THE LIVDET 2015, LIVDET 2019, LIVDET 2021 GREEN BIT DATASETS

<table>
<tr><th>Dataset</th><th colspan="5">Training set</th><th colspan="7">Test set</th></tr>
<tr><td rowspan="2">LivDet 2015 Green Bit DactyScan 26</td><td>Bona fide</td><td>Latex</td><td>WoodGlue</td><td>Gelatine</td><td>Ecoflex</td><td>Bona fide</td><td>Ecoflex</td><td>Gelatine</td><td>Latex</td><td>WoodGlue</td><td>Liquid Ecoflex</td><td>RTV</td></tr>
<tr><td>1000</td><td>250</td><td>250</td><td>250</td><td>250</td><td>1000</td><td>250</td><td>250</td><td>250</td><td>250</td><td>250</td><td>250</td></tr>
<tr><td rowspan="2">LivDet 2019 Green Bit DactyScan 84c</td><td>Bona fide</td><td>WoodGlue</td><td>Ecoflex</td><td colspan="2">BodyDouble</td><td>Bona fide</td><td colspan="2">Mix1</td><td colspan="2">Mix2</td><td colspan="2">Liquid Ecoflex</td></tr>
<tr><td>1000</td><td>400</td><td>400</td><td colspan="2">400</td><td>1020</td><td colspan="2">408</td><td colspan="2">408</td><td colspan="2">408</td></tr>
<tr><td rowspan="2">LivDet 2021 Green Bit DactyScan 84c</td><td>Bona fide</td><td colspan="2">Latex</td><td colspan="2">RProFast</td><td>Bona fide</td><td colspan="2">Mix1</td><td colspan="2">BodyDouble</td><td colspan="2">ElmersGlue</td></tr>
<tr><td>1250</td><td colspan="2">750</td><td colspan="2">750</td><td>2050</td><td colspan="2">820</td><td colspan="2">820</td><td colspan="2">820</td></tr>
</table>

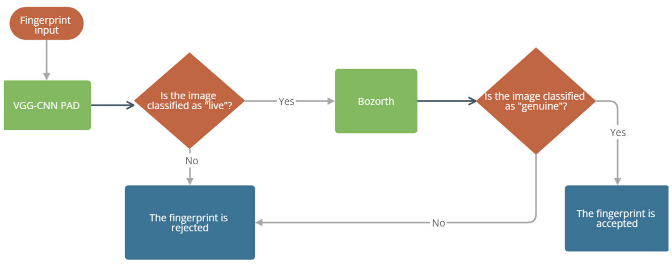


Fig. 11. Schematic of the white-box integrated AFIS consisting of a VGG-CNN PAD and the standard Bozorth3 comparator.

fabricate PAIs and hack a black-box, targeted PAD. The network fine-tuning is necessary to obtain a classifier capable of discriminating a bona fide image from a PA and was done on the LivDet 2015 training set (sec. IV-A). After the white-box protocol allows us to evaluate the effectiveness of the attack from a design point of view, we evaluate the attack from the victim's point of view with a black-box protocol. As black-box PAD and black-box AFIS, we selected eight algorithms submitted to the 2019 and 2021 editions of the Fingerprint Liveness Detection Competition (LivDet) [12], [30]. The algorithms of these two Livdet editions were chosen because these systems are "integrated", i.e. they simultaneously evaluate both the match and the probability of being a PA. In both scenarios, the analyzed PADs work at an acceptance threshold equal to 0.5, i.e. samples with a probability $> 50\%$ are considered bona fide, and samples with a probability $\leq 50\%$ are considered PAs.

The adversarial PAs were obtained from the application of the new Focus Attention techniques (Table III) on 248 latex PAs of the LivDet 2015 test set, originally classified correctly as PA by the network. The adversarial PAIs were acquired through the Green Bit DactyScan 84c scanner.

It is important to notecannot accesstack versions contain incremental changes that lead to the full RFA attack. Therefore, the assessments of such attacks can be read as an ablation study of the Robust Focus Attention Adversarial PA. As a matter of statistical significance, each PAI created was acquired 10 times, slightly modifying each of the 10 acquisitions by varying the angle, the surface fed into the sensor and the applied pressure. The adversarial attacks that have been created physically are those that are successful in the digital domain, thus resulting in a different number for each method and each version of the attack (Table IV). Since the physical PAI crafting procedure and the binarization process might introduce a bias in the PAD score, for all the PA images we also crafted the corresponding physical replica without any adversarial perturbation applied (*re-printed*, see Table III) and

TABLE III

ABBREVIATIONS USED IN THE EXPERIMENTAL EVALUATION TO REFER TO THE ANALYZED VERSIONS OF THE PRESENTATION ATTACKS

Abbreviation	Concept
ADV-PA	Adversarial presentation attack obtained by adding an adversarial perturbation to a digital PA image and subsequent reprinting and acquisition of it.
Re-printed	PA images printed and reacquired with the same original material as the first acquisition
Binarized	PA images are binarized, printed and re-acquired using the same starting material as the original acquisition.
FA_DF	Focus Attention adversarial presentation attack based on DeepFool perturbations.
UFA_DF	Uniform Focus Attention adversarial presentation attack based on DeepFool perturbations.
RFA_DF	Robust Focus Attention adversarial presentation attack based on DeepFool perturbations.
FA_APGD	Focus Attention adversarial presentation attack based on APGD perturbations.
UFA_APGD	Uniform Focus Attention adversarial presentation attack based on APGD perturbations.
RFA_APGD	Robust Focus Attention adversarial presentation attack based on APGD perturbations.

TABLE IV

NUMBER OF SUCCESSFUL ATTACKS IN THE DIGITAL DOMAIN THAT WERE PRINTED FOR THE GENERATION OF THE ADVERSARIAL PAIs. EACH OF THE RESULTING ADVERSARIAL PAIs WAS ACQUIRED TEN TIMES

FA_DF	UFA_DF	RFA_DF	FA_APGD	UFA_APGD	RFA_APGD
147	146	220	210	160	235

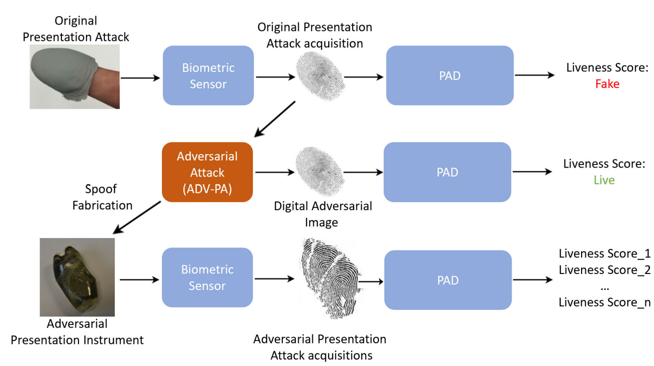


Fig. 12. Process of creation and acquisition of adversarial PAIs. Each PAI is acquired ten times by varying pressure and position on the sensor.

the corresponding binarized replica of the image (*binarized*, see Table III), with the aim of measuring the effect that a simple print and re-acquisition (with or without binarization) has on the AFISs.

1) Performance: We used the following ISO metrics [32], [33]:

- APCER (Attack Presentation Classification Error Rate): the proportion of attack presentations using the same PAI species incorrectly classified as bona fide presentations at the PAD subsystem, that is the rate of presentation attacks classified as bona fide;

- IAPAR (Impostor Attack Presentation Accept Rate): the proportion of PAs using the same PAI species that result in acceptance, that is, the rate of PAs classified as bona fide and mated trial.

This choice is motivated by the need to evaluate the impact of adversarial PAs, whilst the performance on the bona fide, which ISO expresses in terms of BPCER, is unaltered by keeping constant the classification threshold at 0.5. In particular, the average BPCER was 2,58% in LivDet 2019 (page 5 in [30]) and 4,34% in LivDet 2021 (page 6 in [12]).

2) White-Box Test: PAs are acquired with the GreenBit DactyScan 84c scanner and submitted to the VGG-CNN PAD. It is, therefore, a cross-sensor test. Three alternative protocols were used to carry out such accuracy tests:

- *Single attack*: Each PAI acquisition is evaluated on a single sample basis;
- *Multiple attacks*: An ADV-PA with the same PAI is performed ten times. If at least one out of ten is classified as "bona fide", the attack on the PAD is considered successful; if at least one out of ten is classified as "bona fide" and "mated", the attack on the integrated AFIS is successful;
- *Incremental Attack*: this is a sort of ablation study where all versions of the ADV-PAs, including a simple one based on providing a PAI after image binarization, are combined incrementally. If at least one of the PAs is capable of piercing the PAD (for the APCER) or the integrated AFIS (for the IAMPR), the incremental attack is successful.

The white-box protocol allows the analysis of the adversarial PAs from the attacker's point of view.

3) Black-Box Test: The black-box protocol allows the analysis of the adversarial PAs from the point of view of the attacked subject (victim), whilst the previous one represents the attacker's viewpoint in the most effective PAI design. The black-box tests consist of submitting ADV-PAs to completely unknown PAD or integrated AFIS systems (PAD and comparator). This means that the implementation details of neither the PADs nor the comparators are known, nor their integration rule (sequential, in parallel, etc.). Four algorithms participating in LivDet2019 and four algorithms participating in LivDet2021 were used as black-box algorithms. We selected four PADs and four integrated PAD-AFISs, as shown in Table V. It is important to highlight that the black-box algorithms are very different from each other and have been classified into handcrafted, deep-learning and hybrid systems. Furthermore, the LivDet2019 algorithms have been trained on different materials than latex; thus, ADV-PAs are fully "never-seen-before" (Table II). In LivDet2021, the algorithms were trained on PAIs made of latex, and the tests on these algorithms are intra-material.

C. White-Box Results

White-box outcomes impact the attacker's decisions during the PAI's fabrication phase. The attacker aims to fool an AFIS equipped with a PAD with the least effort, that is, with as few attempts as possible.

TABLE V

CHARACTERISTICS OF THE PAD SUBMITTED TO THE TWO EDITIONS OF LIVDET USED AS BLACK-BOX TARGET

	Algorithms	Type	System	Ref.
LivDet2019	PADUnkFv	Handcrafted	PAD	[34]
	Spoof Buster (FSB)	Deep-learning	PAD	[11]
	ZJUT_Det_A	Deep-learning	Integrated	-
	JLWs	Deep-learning	Integrated	-
LivDet2021	contreras	Handcrafted	PAD	[35]
	megvii ensemble	Deep-learning	PAD	-
	hallymMMC	Deep-learning	Integrated	-
	JLWLivDetD	Hybrid	Integrated	-

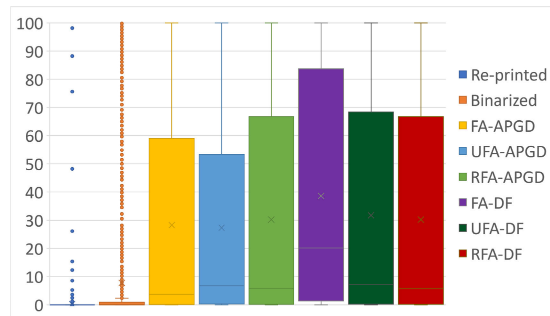


Fig. 13. Boxplots of the VGG-CNN network probabilities after acquiring re-printed original images, printed binarized images and adversarial PAIs.

1) Single Attack: Figure 13 shows the boxplots of the VGG-CNN network probabilities after acquiring the ADV-PAs. The graph includes the boxplot related to the re-printed and binarized samples to evaluate the influence of the re-printing and binarised processes. It is crucial to note that, with the exception of a few outliers, the re-printed image scores are all close to zero. As previously shown in [7], re-printing does not provide effective PAIs. Binarizing the images before re-printing results in a higher percentage of outliers. Also in this case, the percentage of samples with a score far from zero is very low. The attacker cannot, therefore, use these "zero-effort" techniques to reverse the PAD classification. On the other hand, after the application of adversarial manipulations and the printing/acquisition process, a good percentage of attacks exceed the 50% score threshold. This is evident from the results on the PAD, highlighted by the APCER value, and on the integrated system, by the IAMPR value, reported in Table VI. Although the FA_DF method seems to be the most effective (Fig. 13), Table IV shows that the number of created FA_DF samples is lower than RFA_DF and RFA_APGD ones. This is the reason why, although the method allows obtaining very high scores with an average of 40% of "liveness", FA-DF is not the most effective in APCER terms (Table VI): few digital samples have been successfully converted into PAIs resulting in a low number of successful PAs. The most effective attack is RFA_DF with 2.82% of the samples being classified as both bona fide and mated. This was expected as the RFA method is designed to be robust to print by incorporating a dilatation step after binarization.

To show the vulnerability of the white-box PAD to ADV-PAs at different operating points, we have reported in Fig. 14 the APCER vs BPCER DET curve. This figure shows that the previous considerations can be generalized to any operational

TABLE VI

COMPARISON OF THE ADV-PAS ON A WHITE-BOX AFIS WITH THE SINGLE ATTACK PROTOCOL. THE ADV-PAS WERE CREATED STARTING FROM THE PAS CORRECTLY CLASSIFIED AS PAS OF THE ORIGINAL LIVDET2015 TEST SET. EACH ACQUISITION WAS CONSIDERED A SINGLE ATTACK. EACH COLUMN CORRESPONDS TO A PAI SPECIES WHOSE ACRONYMS ARE REPORTED IN TABLE III

	PAI species								
	LivDet2015 Original Test	**Re-Printed**	**Binarized**	**FA_DF**	**UFA_DF**	**RFA_DF**	**FA_APGD**	**UFA_APGD**	**RFA_APG**
APCER	0,00%	0,16%	6,94%	22,10%	18,06%	25,56%	23,31%	16,85%	16,21%
IAPAR	0,00%	0,00%	1,21%	1,77%	1,21%	2,82%	1,01%	1,61%	1,73%

TABLE VII

COMPARISON OF THE ADV-PAS ON A WHITE-BOX AFIS WITH THE MULTIPLE ATTACK PROTOCOL. EACH ATTACK CONSISTS OF 10 ACQUISITIONS: THE ATTACK IS SUCCESSFUL IF ONE OF THE 10 IS CLASSIFIED AS BONA FIDE AND BELONGS TO THE DECLARED USER. EACH COLUMN CORRESPONDS TO A PAI SPECIES WHOSE ACRONYMS ARE REPORTED IN TABLE III

	PAI species							
	Re-Printed	**Binarized**	**FA_DF**	**UFA_DF**	**RFA_DF**	**FA_APGD**	**UFA_APGD**	**RFA_APGD**
APCER	0,81%	22,98%	45,56%	38,71%	55,24%	47,98%	43,55%	41,94%
IAPAR	0,00%	4,84%	4,44%	6,05%	5,65%	4,44%	4,03%	4,44%

TABLE VIII

COMPARISON OF THE ADV-PAS ON A WHITE-BOX AFIS WITH THE INCREMENTAL ATTACK PROTOCOL. THE FIRST COLUMN RELATES ONLY TO BINARIZATION AND RE-PRINTING. THE LAST COLUMN RELATES TO A CONSECUTIVE ATTACK ON THE SENSOR WITH PAIS OBTAINED WITH BINARIZATION AND WITH ALL THE TECHNIQUES FOR CREATING ADV-PAS. THE COLUMNS IN THE MIDDLE ARE RELATED TO THE CONSECUTIVE ATTACK WITH PAIS OBTAINED WITH ONLY BINARIZATION AND ONE OF THE TECHNIQUES FOR CREATING ADV-PAS

	PAI species							
	Bin	**Bin+FA_DF**	**Bin+FA_DF+ UFA_DF**	**Bin+FA_DF+ UFA_DF+RFA_DF**	**Bin+FA_APGD**	**Bin+FA_APGD+ UFA_APGD**	**Bin+FA_APGD+ UFA_APGD+RFA_APGD**	**Bin+All**
APCER	6,94%	25,89%	36,65%	48,59%	27,46%	37,90%	45,44%	64,40%
IAPAR	1,21%	3,27%	3,91%	5,52%	3,67%	4,60%	5,56%	,85%

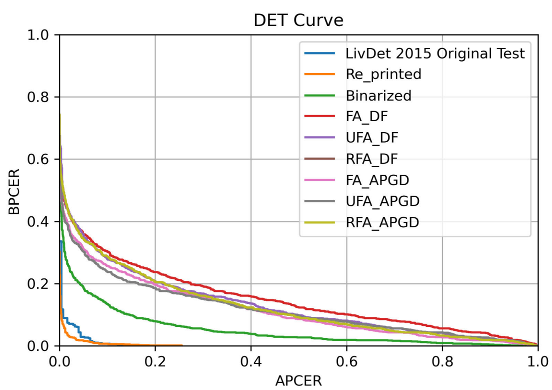


Fig. 14. White-box PAD DET curves comparing the baseline LivDet 2015 samples classification and the ADV-PAs classification.

point: it is, therefore, impossible to simply act on the acceptance threshold to limit the damage of this type of attack.

2) Multiple Attack: This protocol shows the potential danger of an ADV-PA. In fact, having made a successful digital attack, an attacker can exploit it to create multiple PAIs. Additionally, each PAI can be presented multiple times to the sensor. One success out of N trials allows the attacker to access a fingerprint-protected system. Table VII shows the results of this analysis. More than half of the multiple attacks pass the PAD control, and more than 40% are able to cheat the integrated system. Despite the fact that the most effective attacks are RFA_DF for the single PAD system and UFA_DF for the integrated system, it is essential to note that all attacks are successful. Using the proposed focus attention approach, it is possible to manipulate a fingerprint in order to reverse the decision of the PAD while preserving a significant portion of the comparison-sustaining details.

3) Incremental Attack: The worst case for the security of a fingerprint authentication system is an attacker who knows several techniques for making a PAI. In this scenario, the attacker could use all his knowledge to create different PAIs and attack multiple times until the system is broken or evaluate which part of the PAI fabrication algorithm is more effective. The *incremental attack* protocol aims to analyze this possibility, considering multiple attacks consisting of the re-printing of the binarized fingerprints and an adversarial attack, up to the serial use of all the techniques. Combining an adversarial attack with an attempt to print a binarized PA image increases the attacker's chances of success, as demonstrated in Table VIII. This means that the two attacks have a degree of complementarity since fingerprint replicas that were unsuccessful with the adversarial modification are successful with the binarization attack and vice versa. The use of four combined techniques, whether DeepFool-based or APGD-based, exceeds 45% of APCER and 5% of IAPAR. This means that with only four PAIs the attacker has a high chance of cheating the integrated system. Using the seven versions of the attack sequentially allows him/her to reach 64.40% of APCER and almost 7% of IAMPR. It is important to underline that these errors must be added to a normal AFIS performance baseline as these experiments are performed starting from PAs correctly classified [31]. As a result, in a real context, it is necessary to consider the existence of PAs that, without adversarial perturbation, pass through their original characteristics as bona fide.

TABLE IX

RESULTS ON THE TARGET BLACK-BOX PADS SUBMITTED TO THE LIVDET2019 EDITION WITH THE SINGLE ATTACK PROTOCOL. THE 2019 TRAINING SET DOES NOT CONTAIN PAS MADE OF LATEX; THIS EVALUATION IS CROSS-MATERIAL

				PAI species						
	Algorithms	**LivDet2019 Original Test**	**LivDet2021 Original Test**	**Binarized**	**FA_DF**	**UFA_DF**	**RFA_DF**	**FA_APGD**	**UFA_APGD**	**RFA_APGD**
APCER	PADUnkFv	1,55%	9,27%	4,28%	12,86%	55,82%	41,55%	34,9%	40,37%	33,62%
	FSB	0,08%	0,41%	NA	0,20%	6,10%	57,36%	14,62%	6,94%	57,28%
	ZJUT_Det_A	1,14%	1,67%	0,00%	0,07%	1,51%	56,55%	0,48%	1,50%	44,00%
	JLWs	1,14%	1,67%	0,00%	0,07%	1,58%	57,36%	0,57%	1,50%	45,02%
IAPAR	ZJUT_Det_A	2,65%	2,93%	NA	6,92%	1,12%	36,24%	1,41%	9,69%	25,03%
	JLWs	2,60%	1,67%	NA	3,42%	0,52%	31,70%	0,37%	1,10%	21,62%

TABLE X

RESULTS ON THE TARGET BLACK-BOX PADS SUBMITTED TO THE LIVDET2021 EDITION WITH THE SINGLE ATTACK PROTOCOL. THE 2021 TRAINING SET CONTAINS PAS MADE OF LATEX: THIS EVALUATION IS INTRA-MATERIAL

				PAI species						
	Algorithms	**LivDet2019 Original Test**	**LivDet2021 Original Test**	**Binarized**	**FA_DF**	**UFA_DF**	**RFA_DF**	**FA_APGD**	**UFA_APGD**	**RFA_APGD**
APCER	contreras	12,90%	3,94%	0,73%	1,16%	0,41%	7,05%	0,29%	0,31%	8,04%
	megvii_ensemble	0,49%	2,72%	1,58%	0,00%	2,40%	5,18%	3,52%	2,00%	4,04%
	hallymMMC	23,92%	39,17%	32,59%	31,77%	59,52%	54,27%	42,80%	58,62%	44.21%
	JLWLivDetD	6,69%	8,16%	6,62%	2,52%	5,96%	93,09%	24,48%	4,12%	62,04%
IAPAR	JLWLivDetD	6,69%	6,26%	NA	4,44%	2,50%	37,09%	6,07%	1,17%	21,62%

TABLE XI

APCER@1%BPCER AND APCER@10%BPCER FOR LIVDET2019 DETECTORS

			PAI species					
Algorithms	**Error**	**LivDet2019 Original Test**	**FA_DF**	**UFA_DF**	**RFA_DF**	**FA_APGD**	**UFA_APGD**	**RFA_APGD**
PADUnkFv	**APCER@1%BPCER**	0,00%	41,97%	88,42%	77,64%	71,76%	81,44%	71,79%
	APCER@10%BPCER	0,00%	4,63%	35,34%	23,00%	17,38%	18,94%	16,21%
FSB	**APCER@1%BPCER**	0,00%	0,00%	4,73%	73,00%	12,57%	5,63%	52,89%
	APCER@10%BPCER	0,00%	0,00%	0,27%	33,32%	1,00%	0,25%	12,30%
ZJUT_Det_A	**APCER@1%BPCER**	0,00%	0,00%	0,21%	36,05%	0,00%	0,19%	22,68%
	APCER@10%BPCER	0,00%	0,00%	0,00%	3,55%	0,00%	0,00%	1,45%
JLWs	**APCER@1%BPCER**	0,00%	0,00%	0,21%	37,00%	0,00%	0,19%	23,19%
	APCER@10%BPCER	0,00%	0,00%	0,00%	3,77%	0,00%	0,00%	1,87%

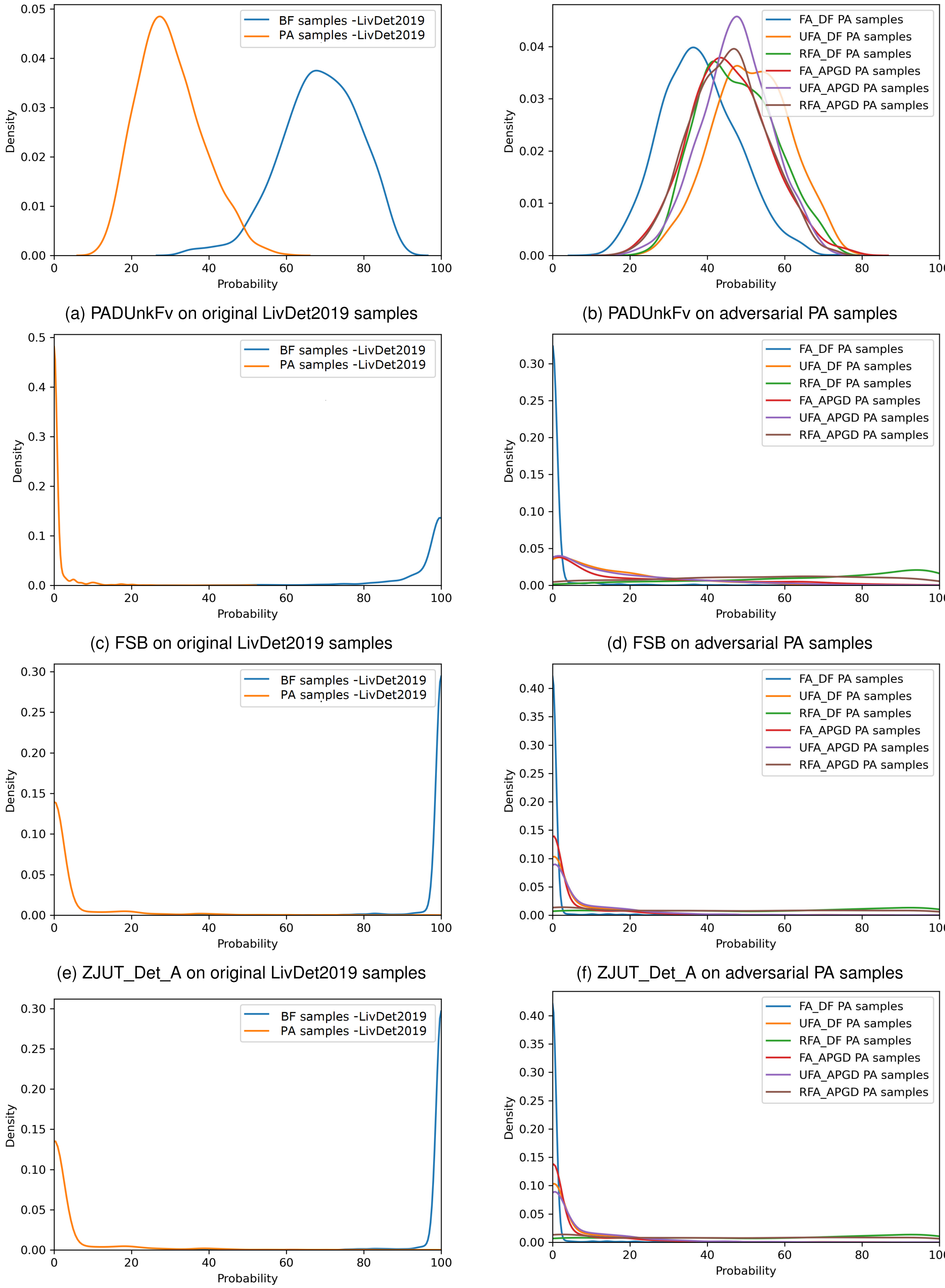
TABLE XII

APCER@1%BPCER AND APCER@10%BPCER FOR LIVDET2021 DETECTORS

			PAI species					
Algorithms	**Error**	**LivDet2019 Original Test**	**FA_DF**	**UFA_DF**	**RFA_DF**	**FA_APGD**	**UFA_APGD**	**RFA_APGD**
contreras	**APCER@1%BPCER**	0,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	APCER@10%BPCER	0,00%	1,00%	0,41%	7,05%	0,29%	0,31%	8,04%
megvii_ensemble	**APCER@1%BPCER**	0,00%	0,00%	0,00%	0,27%	0,00%	0,00%	0,64%
	APCER@10%BPCER	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
hallymMMC	**APCER@1%BPCER**	0,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	APCER@10%BPCER	0,00%	99,86%	100,00%	100,00%	99,90%	99,81%	99,91%
JLWLivDetD	**APCER@1%BPCER**	0,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	APCER@10%BPCER	0,00%	0,20%	0,21%	78,05%	6,24%	0,00%	32,21%

D. Black-Box Results

The main objective of this work is to evaluate whether it is possible to carry out an ADV-PA in completely black-box mode. In this case, the attacker does not know the biometric recognition system they want to attack. This also allows us to evaluate how much the previous analysis leads the attacker to provide general-purpose PAIs. On the other hand, the following results allow us to evaluate ADV-PAs from the victim's point of view in terms of potential damage needing counteraction. For a fair comparison, the error rates resulting from the different versions of ADV-PAs were compared with the results of the methods on the original LivDet2019 and LivDet2021 datasets. Table IX displays the outcomes generated by the LivDet2019 algorithms. In the case of integrated systems, the greatest risk associated with the attack is evidenced by high APCERs and IAPARs. In this case, the attack is effective both from the point of view of the PAD and the comparator. From this point of view, the RFA_DF attack is the most dangerous of all. For PADs without integration, we can instead evaluate only how many PAs are classified as bona fide samples. From this point of view, we can highlight the difference between handcrafted and deep-learning methods. While the latter seems to be



(a) PADUnkFv on original LivDet2019 samples

(b) PADUnkFv on adversarial PA samples

(c) FSB on original LivDet2019 samples

(d) FSB on adversarial PA samples

(e) ZJUT_Det_A on original LivDet2019 samples

(f) ZJUT_Det_A on adversarial PA samples

(g) JLWs on original LivDet2019 samples

(h) JLWs on adversarial PA samples

Fig. 15. Analysis of the distribution of the output scores of the analyzed black-box PADs.

exposed to RFA_DF and RFA_APGD PAIs, the former suffers from all PAIs. In particular, UFA-based PAIs include uniform perturbations on the ROI, and RFA-based PAIs include all the pre-processing steps, that is, binarization and dilation, applied to make the adversarial modifications robust to the printing process.

Table X shows the results of the LivDet2021 detectors. In this case, the handcrafted algorithm appears to be the less vulnerable, especially when comparing the error rates of the analogous handcrafted algorithm in Table IX. However, it should be noted that RFA-based PAIs are much more effective: APCER is higher than the baseline one (fourth column of Table X). Hybrid and deep-learning methods, on the other hand, are vulnerable to ADV-PAs. For example, the 93.09% of the PAs obtained with the RFA_DF method cheated the JLWLivDetD PAD and the 37.09% are classified by the integrated system as bona fide and mated (last row of Table X). The substantial differences among systems show that the vulnerability to ADV-PAs is highly linked to the type of PAD and comparator implemented. However, the attacks represent a potential danger for all of them. In the worst cases, IAMPR values are above 30%, very high for an authentication system.

We remark that the results obtained are related to an acceptance threshold set at 0.5 in both PAI's design and attack phases. In this scenario, the victim could try counteracting by modifying the classification threshold, making it more stringent. Thus, we calculated the APCER when BPCER=1% and BPCER=10% and reported such values in Tables XI- XII.

Moreover, we analyzed the outcomes of the networks, shown in Fig.15. These values correspond to the input image's probability of being *bona fide*. In particular, each row corresponds to the outputs of a particular PAD: on the left, the ones of LivDet2019 original samples, and on the right, the ones of adversarial PAs. The RFA-based PAs curves (green and brown curves) show in all cases a portion of samples with very high scores, **i.e.**, high probabilities of being classified as *bona fide*. Consequently, the acceptance threshold increase would impact the classification of bona fide samples. Since the score distributions for both adversarial attacks and bona fide presentations overlap, setting a higher threshold is not effective. This is confirmed by values reported in Tables XI-XII.

V. Discussion and Conclusion

Digital adversarial attacks are effective against modern fingerprint authentication systems even when protected with PAD. However, they presuppose access to internal modules of the system and are therefore unrealistic. This risk becomes concrete when the modified fingerprint is brought into the physical domain and submitted to the sensor, generating an adversarial presentation attack. However, these kinds of attacks were only tested on white-box classifiers, namely, the same systems adopted to fabricate the PAIs. Results reported so far were partial and did not allow for assessing if adversarial presentation attacks were effective when conducted on systems whose nothing is known. To better investigate this perspective, in this paper, we followed all the phases of the attack, from the perturbation algorithm to the realization of the PAI. The selection of an appropriate white-box PAD and AFIS simulated the attacker's tools for the PAIs fabrication. Finally, we provided a statistically significant set of PAIs to carry out adversarial presentation attacks on fully black-box integrated fingerprint recognition systems. With regard to the PAI fabrication, we proposed a focus attention mechanism able to align the perturbation and the ROI of the white-box classifier by introducing a novel multi-stage process along with saliency-aware guidance feedback, utilizing Explainable AI (XAI) methods. We showed that this made the attack more effective and allowed us to maintain effective adversarial perturbations after the printing process. This was confirmed in the design phase by a set of experiments under the white-box protocol and in the attack phase under the black-box protocol. In particular, the latter set of experiments was carried out by adopting state-of-the-art PADs and integrated AFISs from LivDet 2019 and LivDet 2021 competitions and represents the first comprehensive analysis of the use of presentation adversarial attacks in a realistic attack simulation on integrated fingerprint recognition systems.

With the white-box protocol, we showed the effectiveness of the focus attention mechanism proposed, highlighting that an attacker can rely on the complementarity of different techniques to generate more PAIs and increase her/his probability of success.

From the victim's point of view, the black-box protocol proved the need to protect systems from adversarial presentation attacks. In fact, the results showed that it is possible to cheat completely black-box AFISs with PAD modules obtaining APCER greater than 50% and IAMPR greater than 30%. Obviously, knowing that these attacks are possible enables AFIS designers to safeguard their systems. A trivial solution could be to train the PAD even on adversarial PA samples. However, this is both costly and time-consuming; it requires high knowledge of fingerprint replication and adversarial techniques, leading to a real "vicious circle" for attackers and defenders. Other strategies might be proposed to exploit the common characteristics of standard and adversarial PAs. Alternatively, the two issues might be addressed in a blended manner. For example, since adversarial perturbations have been shown to modify the frequencies of the resulting image spectrum [36], [37], an analysis in the frequency domain could be exploited for the detection of adversarial presentation attacks under the hypothesis that this frequency is kept along the PAI fabrication process.

As a matter of fact, a thorough analysis must be conducted to assess if this is true and to what extent; this is the next step in our research pathway.

References

[1] A. Roy, N. Memon, and A. Ross, "MasterPrint: Exploring the vulnerability of partial fingerprint-based authentication systems," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 9, pp. 2013–2025, Sep. 2017.

[2] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artificial 'gummy' fingers on fingerprint systems," *Datenschutz und Datensicherheit*, vol. 26, no. 8, 2002, Art. no. 462719.

[3] S. Marcel, M. S. Nixon, J. Fiérrez, and N. W. D. Evans, *Handbook of Biometric Anti-Spoofing—Presentation Attack Detection* (Advances in Computer Vision and Pattern Recognition), 2nd ed. Cham, Switzerland: Springer, 2019.

[4] T. Chugh and A. K. Jain, "Fingerprint presentation attack detection: Generalization and efficiency," in *Proc. Int. Conf. Biometrics (ICB)*, 2019, pp. 1–8.

[5] S. Marrone and C. Sansone, "Adversarial perturbations against fingerprint based authentication systems," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–6.

[6] B. Zhang, B. Tondi, and M. Barni, “Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability,” *Comput. Vis. Image Understand.*, vols. 197–198, Aug. 2020, Art. no. 102988.
[7] S. Marrone, R. Casula, G. Orrù, G. Marcialis, and C. Sansone, “Fingerprint adversarial presentation attack in the physical domain,” in *Pattern Recognition*. Cham, Switzerland: Springer, 2021, pp. 530–543.
[8] B. Biggio, G. fumera, P. Russu, L. Didaci, and F. Roli, “Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective,” *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 31–41, Sep. 2015.
[9] Z. Akhtar, C. Micheloni, and G. L. Foresti, “Biometric liveness detection: Challenges and research opportunities,” *IEEE Secur. Privacy*, vol. 13, no. 5, pp. 63–72, Sep. 2015.
[10] S. A. C. Schuckers, “Spoofing and anti-spoofing measures,” *Inf. Secur. Tech. Rep.*, vol. 7, no. 4, pp. 56–62, Dec. 2002.
[11] T. Chugh, K. Cao, and A. K. Jain, “Fingerprint spoof buster: Use of minutiae-centered patches,” *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2190–2202, Sep. 2018.
[12] R. Casula et al., “LivDet 2021 fingerprint liveness detection competition–into the unknown,” in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–6.
[13] C. Szegedy et al., “Intriguing properties of neural networks,” 2013, *arXiv:1312.6199*.
[14] B. Biggio et al., “Evasion attacks against machine learning at test time,” in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2013, pp. 387–402.
[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” 2017, *arXiv:1706.06083*.
[16] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
[17] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Mar. 2016, pp. 372–387.
[18] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: A simple and accurate method to fool deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.
[19] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2206–2216.
[20] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.
[21] M. A. A. Milton, “Evaluation of momentum diverse input iterative fast gradient sign method (M-DI2-FGSM) based attack method on MCS 2018 adversarial attacks on black box face recognition system,” 2018, *arXiv:1806.08970*.
[22] C. Xie et al., “Improving transferability of adversarial examples with input diversity,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2019, pp. 2730–2739.
[23] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, “Towards transferable targeted attack,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2020, pp. 641–649.
[24] S. Marrone and C. Sansone, “On the transferability of adversarial perturbation attacks against fingerprint based authentication systems,” *Pattern Recognit. Lett.*, vol. 152, pp. 253–259, Oct. 2021.
[25] R. Casula, G. Orrù, D. Angioni, X. Feng, G. L. Marcialis, and F. Roli, “Are spoofs from latent fingerprints a real threat for the best state-of-art liveness detectors?” in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3412–3418.
[26] X. Zhang, F. T. Chan, and S. Mahadevan, “Explainable machine learning in image classification models: An uncertainty quantification perspective,” *Knowl.-Based Syst.*, vol. 243, May 2022, Art. no. 108418.
[27] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision–ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 818–833.
[28] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, “Fingerprint liveness detection using convolutional neural networks,” *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 6, pp. 1206–1213, Jun. 2016.
[29] V. Mura, L. Ghiani, G. L. Marcialis, F. Roli, D. A. Yambay, and S. A. Schuckers, “LivDet 2015 fingerprint liveness detection competition 2015,” in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2015, pp. 1–6.
[30] G. Orrù et al., “LivDet in action–fingerprint liveness detection competition 2019,” in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–6.
[31] M. Micheletto, G. L. Marcialis, G. Orrù, and F. Roli, “Fingerprint recognition with embedded presentation attacks detection: Are we ready?” *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5338–5351, 2021.
[32] *Information Technology—Biometric Presentation Attack Detection—Part 3: Testing and Reporting*, Standard ISO/IEC 30107-3:2023, 2023.
[33] *Information Technology—Vocabulary—Part 37: Biometrics*, Standard ISO/IEC 2382-37:2022, 2022.
[34] L. J. González-Soler, M. Gomez-Barrero, L. Chang, A. Pérez-Suárez, and C. Busch, “Fingerprint presentation attack detection based on local features encoding for unknown attacks,” *IEEE Access*, vol. 9, pp. 5806–5820, 2021.
[35] R. C. Contreras et al., “A new multi-filter framework with statistical dense sift descriptor for spoofing detection in fingerprint authentication systems,” in *Artificial Intelligence and Soft Computing*. Cham, Switzerland: Springer, 2021, pp. 442–455.
[36] Y. Zhou, X. Hu, J. Han, L. Wang, and S. Duan, “High frequency patterns play a key role in the generation of adversarial examples,” *Neurocomputing*, vol. 459, pp. 131–141, Oct. 2021.
[37] N. Yu, L. Davis, and M. Fritz, “Attributing fake images to GANs: Learning and analyzing GAN fingerprints,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7555–7565.

Open Access funding provided by ‘Università degli Studi di Cagliari’ within the CRUI CARE Agreement