The Alan Turing Institute

Technical Briefing

Attacks Against Face Recognition Systems: A State-of-the-art Review

January 2023

Authors

Professor Carsten Maple, Turing Fellow, Project Principal Investigator, and Professor of Cyber Systems Engineering with Institute partner University of Warwick

Dr Gregory Epiphaniou, Associate Professor in Security Engineering, University of Warwick

Dr Roberto Leyva, Research Associate, University of Warwick

This Technical Briefing is published by The Turing's Trustworthy Digital Infrastructure for Identity Systems project.

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation [INV-001309]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript.

The Institute is named in honour of Alan Turing, whose pioneering work in theoretical and applied mathematics, engineering and computing is considered to have laid the foundations for modern-day data science and artificial intelligence. It was established in 2015 by five founding universities and became the United Kingdom's (UK) National Institute for Data Science and Artificial Intelligence. Today, the Turing brings together academics from 13 of the UK's leading universities and hosts visiting fellows and researchers from many international centres of academic excellence. The Turing also liaises with public bodies and is supported by collaborations with major organisations.

The Alan Turing Institute

British Library

96 Euston Road

London

NW1 2DB

Table of Contents

1 Purpose

As the United Kingdom's national institute for data science, The Alan Turing Institute is driving research into how digital identity systems are evolving to underpin a changing world, including their impact on people and communities to elevate the requirements for assuring trustworthy outcomes. This Technical Briefing presents a focused area of work to elevate understanding of how systems vulnerabilities and attacks are evolving with growing adoption of facial recognition technologies in these systems. It represents a foundational step in the Institute's work to advance knowledge of the threats and associated risks specific to identity systems and inform efforts to model them. It forms the basis of specific work to define a New Authentication Reference Architecture (ARA) and Taxonomy for Modelling Threats published alongside this Briefing and contributes to a body of resources and guidance developed for and in consultation with governments, humanitarian organisations and the industry stakeholders that are advancing digital identity systems.

2 Executive summary

There is a growing trend to adopt Facial Recognition Systems in many identification and user verification or authentication processes, including within national foundation identity programmes, online commerce, border controls, banking and online market places that increasingly underpin modern economies. As governments and society develop new reliance on digital identity technologies, such as facial recognition, it is important that we recognise measures to be taken that can attest to their trustworthiness.

A big challenge before us lays in understanding how vulnerabilities and risks are changing so that the underlying systems and processes can be designed to manage them. There are many factors to consider here. The incentives to misuse, commit fraud, breach or manipulate these systems are growing with their scope. Tensions arise from competing goals, between security and privacy for example, as systems evolve to facilitate transparency for users, and enforce limitations on data use across the system. The protocols, architectures and technical controls for data collection, processing, and release that are designed into the system are critical.

Turing researchers are examining these challenges through a broad lens provided by six pillars of trustworthiness—security, privacy, robustness, ethics, reliability, and resiliency—to define aspects that determine reliability of access to resources and services, the appropriateness of their use and the sustainability of design in terms of the technology, social and economic environments in which they operate.

Here we present an extensive review of the most persistent threats towards facial recognition systems deploying a systematic reviewing process, known as PRISMA. This process was applied to analyse publications across five major publishers and online resources, covering records from 2018 – 2021 of the most persistent threats to face recognition systems. We created a web crawler to retrieve specific keywords to build a database, generating the keyword set from surveys and well-known attacks dating back to 2012, in order to identify threats present in the most persistent attacks and present a taxonomy of them. Our PRISMA comprises recently discovered attacks along with their frequency in the reported literature, revealing insightful information, for instance, that some of the most severe threats do not require a particularly high level of technical sophistication from the attacker's background.

3 Extended Literature Review

3.1 Acronyms

APT Advanced Persistent Threats.

ARA Authentication Reference Architecture

CaaS Cybercrime as a Service.

FRS Face Recognition Systems.

GAN Generative Adversarial Network.

IoC Indicator of Compromise

NNIST National Institute of Standards and Technology.

OSI Open Systems Interconnection.

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

RaaS Ransomware as a Service.

RGB RedGreenBlue

ROC Receiver Operating Characteristics

SPN Sensor Pattern Noise.

3.2 PRISMA

To elaborate our threat models, we carried out an evidenced-based PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) over four major publishers and two online sources, Fig.1 illustrates this process. We systematically searched for threats previously described in the literature from 2012– 2021 and used them as keywords (104) to run on the publishers' engines from 2018 – 2021. We then created a crawler to conduct the search, limiting the response to 100 records per keyword. The tool gave us back approximately 22,000 results in the first run. We proceeded to remove duplicates leaving nearly 16,000 records. Next, we proceeded to filter the title and description, excluding other subjects of study, e.g., power grid, antenna, that appear very frequently, which reduced the records to c.a. 3500 records. Next, we manually inspected non-related subjects seeking title (first) followed by the abstract (second) to leave 393 records. Finally, we selected 233 papers based on the abstract.

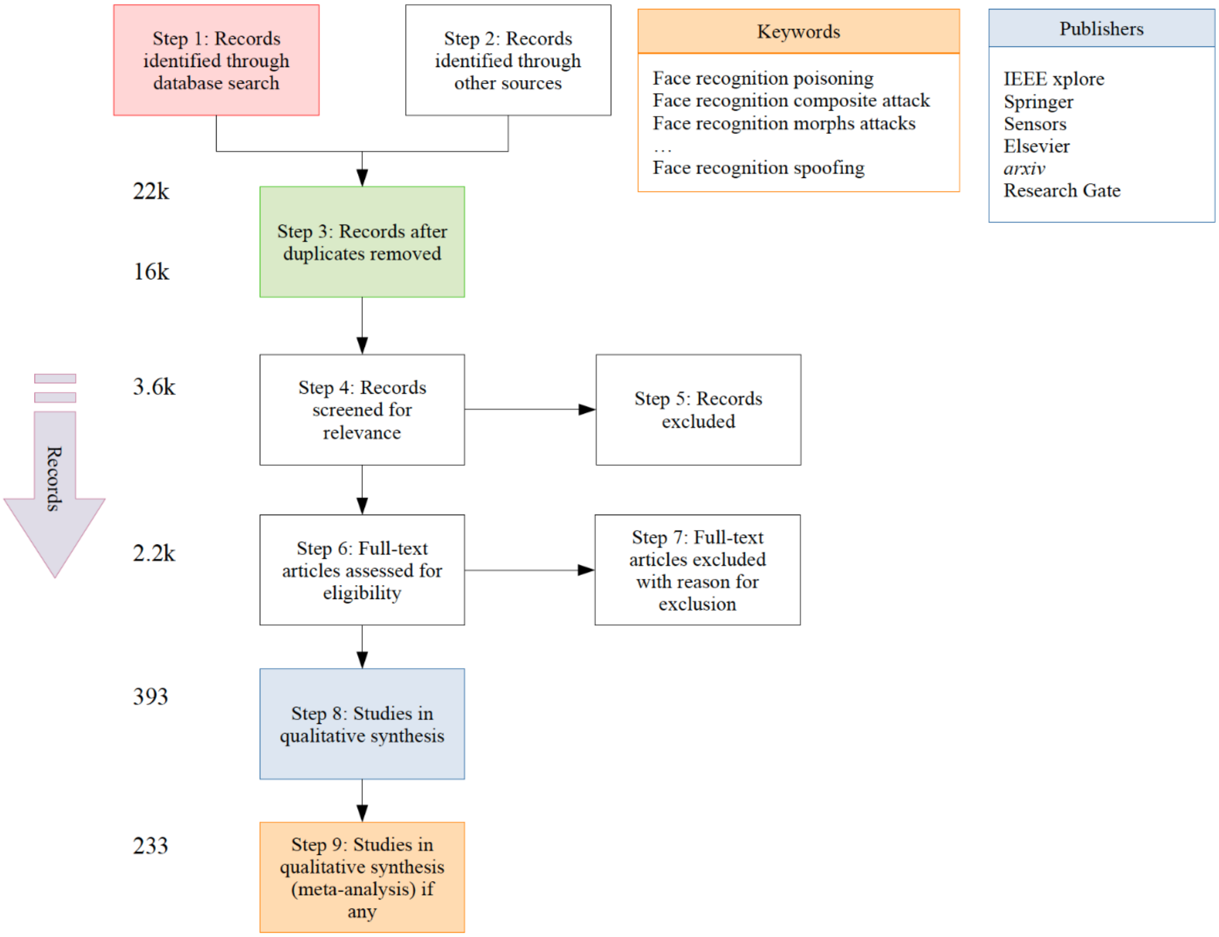

Step 1: Records identified through database search
Step 2: Records identified through other sources
Keywords
Face recognition poisoning
Face recognition composite attack
Face recognition morphs attacks
…
Face recognition spoofing
Publishers
IEEE xplore
Springer
Sensors
Elsevier
arxiv
Research Gate
22k
16k
Step 3: Records after duplicates removed
3.6k
Step 4: Records screened for relevance
Step 5: Records excluded
Records
2.2k
Step 6: Full-text articles assessed for eligibility
Step 7: Full-text articles excluded with reason for exclusion
393
Step 8: Studies in qualitative synthesis
233
Step 9: Studies in qualitative synthesis (meta-analysis) if any

Fig 1: PRISMA approach for literature review to identify the records.

3.3 Threats Towards Face Recognition Systems

Previous work mainly focuses on modelling threats that specifically aim towards presentation and model threats, which are by far the most persistent and therefore most important threats. We can group threat models into five categories depending on their nature. Presentation group: This group concerns those attacks targeting the system's input, where a fabricated or tampered sample is given as a bonafide sample. It is important to notice that we may have two different purposes, to impersonate someone or gain access to the system no matter who. Template group: this group regards attacks targeting processed data from the input samples, thus aiming the data to be stored and used for authentication purposes. Model group: In this group, the attacks aim to corrupt the model's operation; this can classify one subject as another or completely avoid the system's authentication process. Hardware group: this group regards threats from the electronics involved in the sample acquisition process. Transmission/Storage group: this group concerns threats aimed at the data handling mainly involving data spoofing (impersonation) and biometric injection attacks (fraudulent use of synthetic or genuine biometrics). However, it is essential to mention that we may have a vast catalogue of threats falling into this category that are also relevant to other user authentication systems. Thus, this report bounds its scope for those reportedly being used for face recognition purposes. We can group threats based on their nature and targets. Fig.2 illustrates the five threat groups we identified.

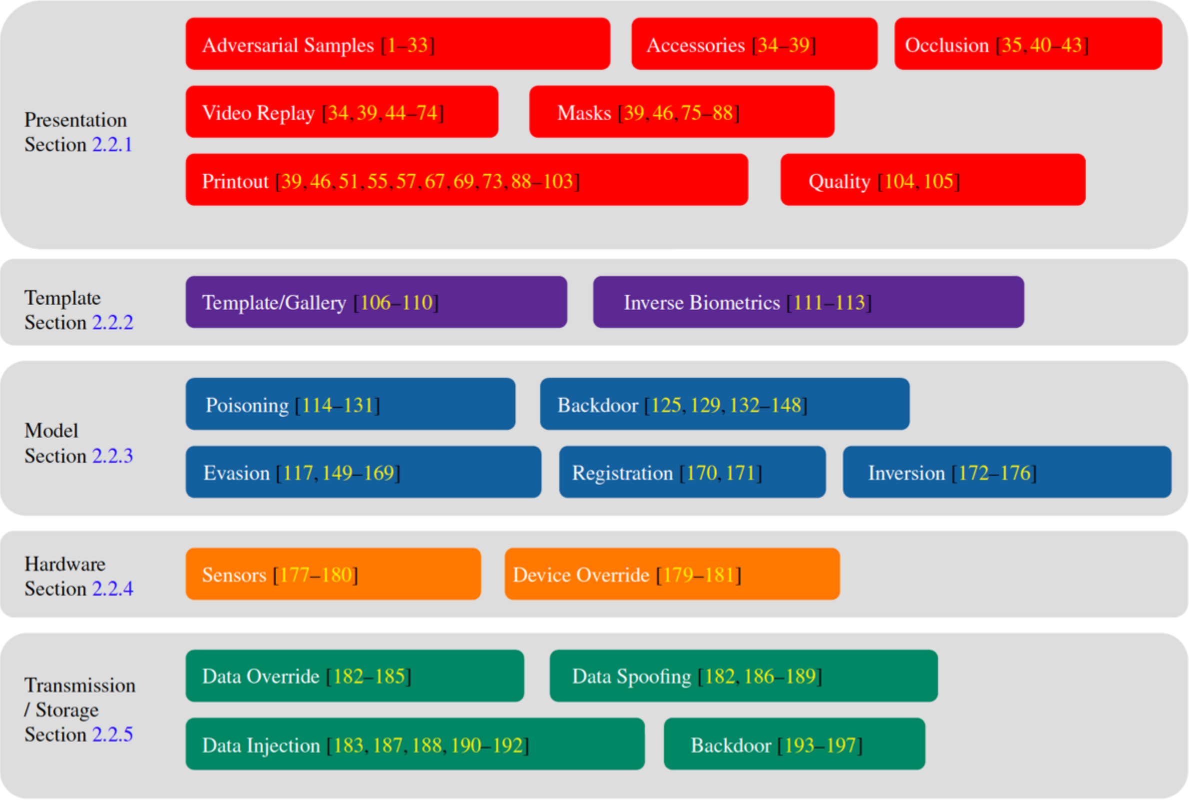


Fig 2: facial recognition system *attack groups. The diagram depicts the threat groups for user-identification systems based for face recognition.*

4 Types of Attack

4.1 Presentation

Presentation threats involve producing a fabricated/tampered sample to bypass the system's authentication. The threat vector is to identify users. Once the victim is identified it is necessary to obtain samples from him/her. To this end, social media platforms and public websites are plentiful resources, as they comprise free available samples sometimes without the victim's knowledge [198]. Although more difficult, samples can also be obtained live in person without the victim's consent: For example, pictures can be taken in public places without being a privacy violation. The next step is to generate the image replica which will be used to bypass the system. These images will eventually lead to impersonating a specific subject (spear impersonation) or any person that uses the FRS (random impersonation). We detail the image generation next and assume that the attacker can identify the victims and retrieve face samples from them.

Presentation threats comprise three broad categories where the attacker would produce a tampered sample or use some artifact to impersonate someone else. Note that the latter does not necessarily mean the attacker tries being specific, just being authorised. i.e., spear impersonation. The third category concerns concealing, where the objective is to cancel the system functionality. This may be to avoid detection in a surveillance scenario or make digital forensics difficult in an event of a crime.

4.2 Printouts

The simplest attack is the printout. This attack requires only a sample image of the victim and a high-quality printer to create the replica. As a consequence, high-definition printouts [51,93] present a problem for FRS. Fig.3 shows examples of printouts used to bypass FRS. Intuitively, these attacks require little level of sophistication once the victim has been identified.

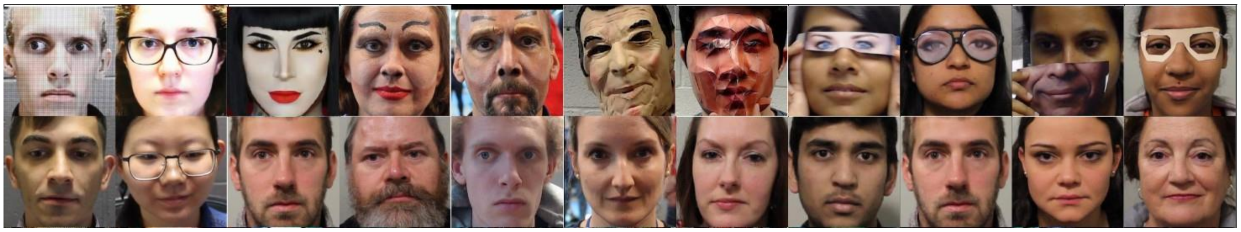


Fig 3: *Printout samples to perform impersonation attack. First row printout, second row live targets.*

4.3 Video Replay

Some authentication methods imply detecting whether the face is actually alive to avoid the printout attack, these are known as liveness detection methods [77]. Because some spoofing detection methods are based on motion detection [66–68], these require computing motion characteristics, e.g., optical flow, and tracking facial landmarks to determine whether the subject is alive. Given this, presentation attacks have been perpetrated using videos instead of images, Ming *et al.* [34] and Sepas-Purnapatra *et al.* [199,200] detail attacks by means of video replay. The attack implies having a video of the intended victim and providing it as a sample to the FRS. The attack then requires an electronic device to be performed. Elementary software, such as Kazam Screencaster or camfecting, where a device camera is surreptitiously turned on, can produce high-quality video posing a threat to FRS via video replay attacks [56–60]. Fig.4 shows some examples of the video replay attack. Some electronic devices, e.g. iPads produce samples almost imperceptibly. Compared to printouts, video replays are more sophisticated, as they introduce intrinsic dynamic information to mimic liveness, e.g., eye blinking, mouth movements, and changes in facial expressions [34].

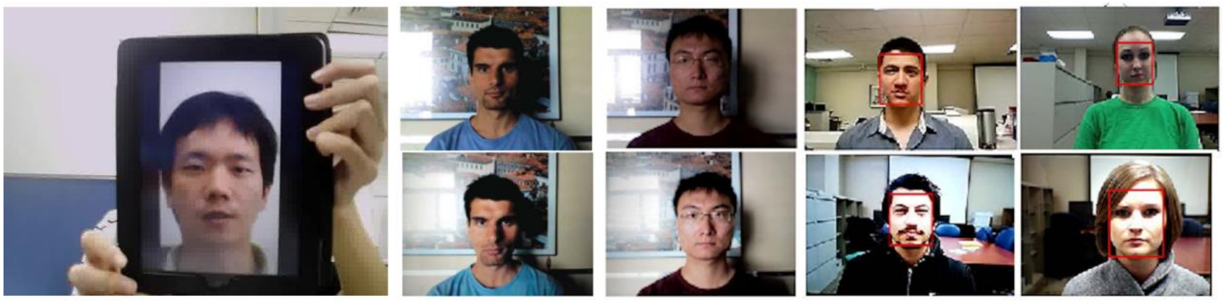


Fig 4: *Video replay samples to perform impersonation attack. First row live samples, second row device's video.*

4.4 Accessories

We find two different purposes for this attack vector. One is aiming for spear impersonation and the other is avoiding the FRS functionality. The latter is mainly related to avoiding detection of a specific person identification, e.g. in the case of a manhunt. For spear impersonation, we find that makeup can create very accurate face replicas. Fig. 5 (bottom) shows samples of the generated faces [36]. Where it is difficult to distinguish the real person from the phony. This attack requires a sophisticated artist to perform, posing a high challenge to liveness detection methods [38]. On the other hand, we find the accessories for disguising purposes. These attacks could be elaborated using accessories, e.g., sunglasses, make-up, fake beards [34,36]. Intuitively none of these requires a sophisticated attacker to produce the samples. Fig. **5** shows concealer presentation attacks via accessories. Although face disguise recognition is well addressed in the literature, it is difficult to recognise data subjects who present themselves with a disguise. The central idea is to conceal the identity by changing the subject's appearance. This subject will not be identifiable in an open scenario, e.g. public place. Accessories attacks can be performed easily, especially in automatic access control applications such as entry and exit control in football stadiums, and restricted areas. In the same line, a recently discovered attack [92] requires placing a sticker on the top of the head to completely stop the system from correctly classifying subjects. The sticker is a printout of a deformed face via off-plane transformations, which in a nutshell is a cascade of affine transformations (a linear mapping method that preserves points, straight lines, and planes). These are the attack vector core, where the FRS considers the patch as part of the head and collapses the classification capabilities. This attack presents a serious threat to surveillance systems, e.g. the ones used in border control as they aim to spot criminals in an online automated fashion.

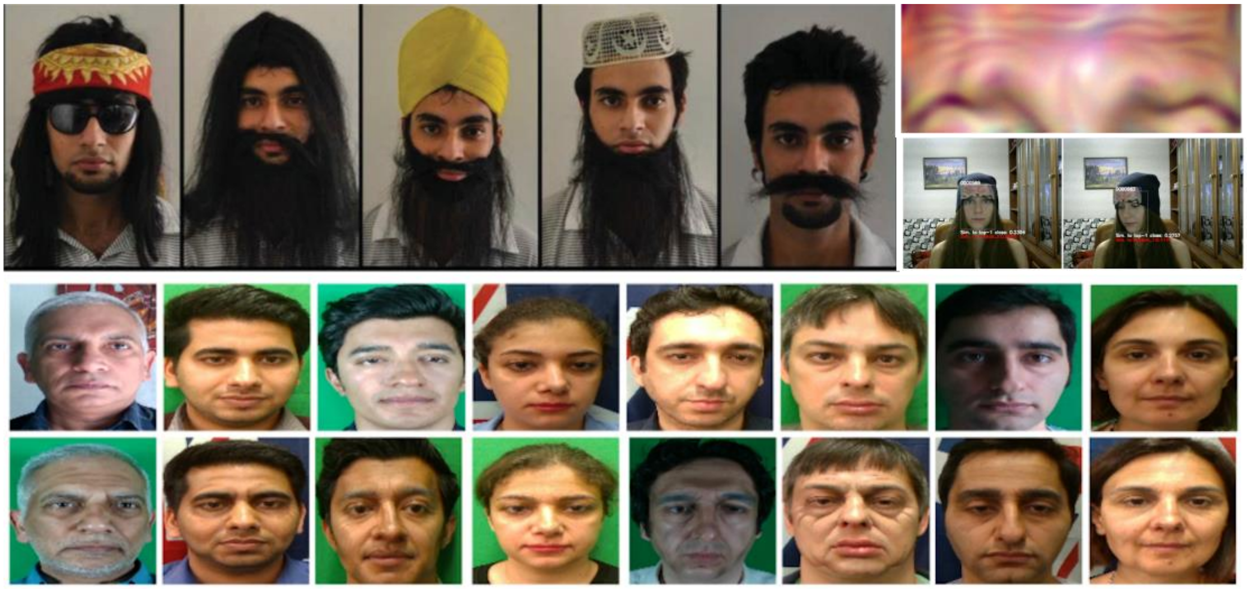


Fig 5: *Accessories samples to perform impersonation/evasion attack [36] (left). Adversarial patch on the forehead [92] (right). Makeup samples [37] (bottom).*

4.5 Occlusion

Occlusions are also threats [40,41] that exploit the fact that people may have very similar facial features. In such scenarios, patches or areas of a face make it difficult to distinguish the subjects. Two people may look very similar when wearing black glasses for instance. It is possible to train a classifier to identify subjects based on carefully selected patches and achieve state-of-the- art accuracy in user identification [41]. Fig.6 shows samples from two methods and the activation areas. The fake samples (blue areas) show the areas needed to perform the classification and detected by the system as fake, while the real examples require the whole face to perform the same task (red areas). The activation map shows the average of the method areas needed to perform the classification correctly, revealing that hair, nose, and mouth are the most important parts to do the task. Subsequently, the attacker could perform impersonation by carefully alternating those areas to create a new face. This reveals that existing models need only partial information from face landmarks to perform the prediction, making the occlusion attack extremely feasible.

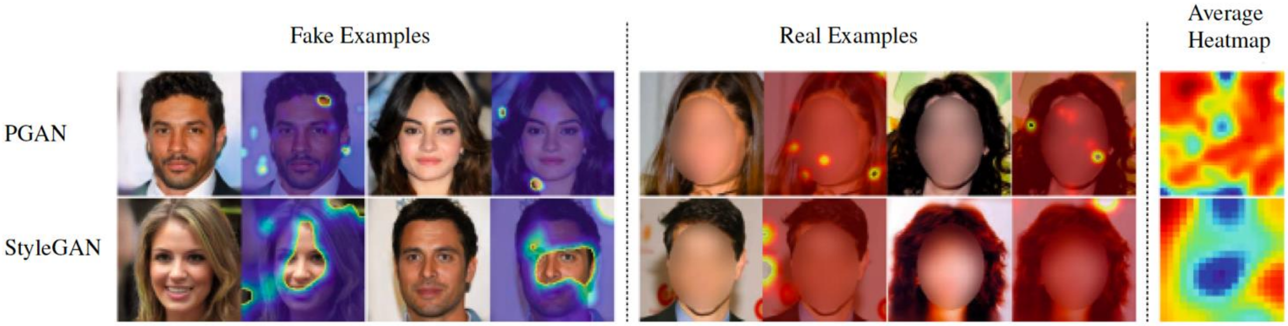


Fig 6: *Occlusion attacks, some face landmarks can be used to classify subjects no matter the rest of the face [41]* .

4.6 Masks

High-quality masks [46,75,76] pose serious threats to FRS. We find three main groups in this category. The first is generating masks using wax. These are extremely similar to the reference subject however, liveness-based methods are not easy to fool with this technology [78]. Wax techniques require sophisticated artistic skills. The second group is silicone masks. These have the potential capability of adding dynamic facial characteristics, particularly regarding eyes and mouth. A presentation attack using them is more complex than via still masks [86]. These masks also require elaborate technical and artistic skills. The last group is 3D-masks. This is an emerging threat as recent technological advances have made 3D printing accessible. A realistic 3D mask resembles real skin textures [79], including for example wrinkles, presenting a significant challenge to the prevention of such an attack, and often a significant number of approaches are used to detect it. As in the case of wax masks, the 3D mask attack does not counter liveness-based methods. [89]. Overall with hyper-realistic printing we can argue that printing advances must always be considered as a potential security threat.

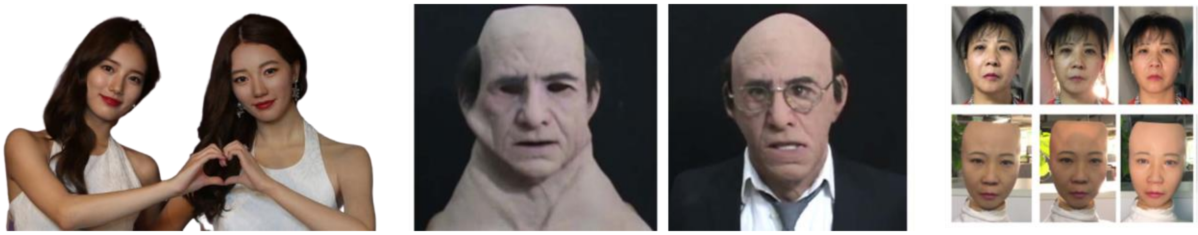


Fig 7: *Masks samples for impersonation/evasion. Wax sample replicating not only the face but the whole subject [78] (left). Silicone mask to encompass the whole head [86] (middle). 3D masks creates the face only [80] (right).*

4.7 Quality

Low quality and distorted samples also pose threats to facial recognition systems. The model's accuracy degrades in the presence of challenging appearance variations, e.g., background, illumination, diverse spoofing materials, and low image quality [104,105]. Wang *et al.* [104] propose a strategy using GANs as a domain-transfer-module to convert an image to a depth map. The live face image is transferred to a depth map where the facial details are preserved, whereas spoofed faces are transferred to a black plain image. The method heavily depends on the latent variables that can effectively represent the depth information comprised in the original RGB space. The image quality is a cornerstone of the method. Fig.8 shows samples from a printout attack and the activated regions to detect the spoofing. High-resolution samples fool RGB liveness detection strategies using for example remote photoplethysmography [105]. Thus, a potential attack is flooding a dataset with low-quality reference data to open the door to adversarial attacks which heavily rely on good quality data. The attack then takes place at the registration time, weakening the reference data.

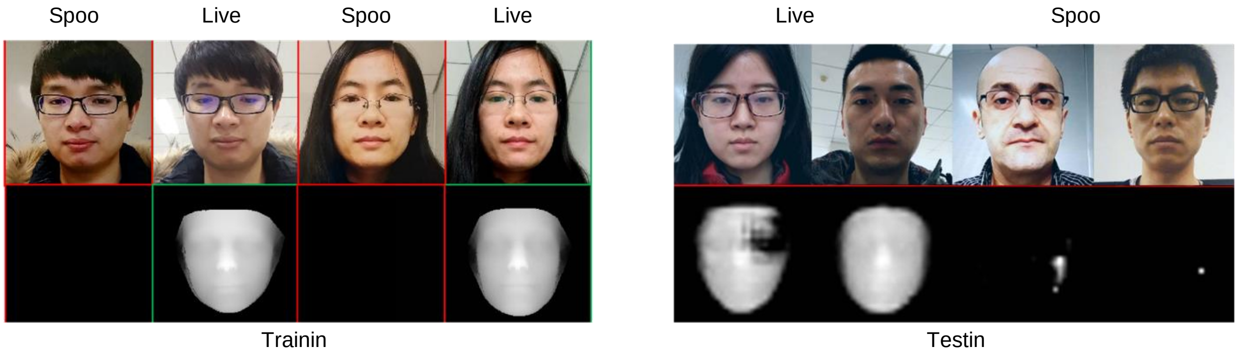


Fig 8: Spoofing and live samples, the GAN-based method to detect the attack relies on high-resolution RGB samples.

More elaborated threats are adversarial biometric sample attacks. These deploy machine learning techniques to fabricate samples. The attacker uses a model to produce similar biometric samples from a bonafide one. For this scenario, we find two main branches. The first are aligned morphs, a technique that requires mixing two or more subjects to create a single face using elementary linear maps. The second is generative morphs, a technique to create samples on a latent feature space embedding many candidates. Morph attacks require the attacker to mix the collaborators' or victims' samples with an impostor's one [2,19,19, 30,201]. The technique is based on combinations of the pixels from the two images to create the morph. It is necessary for the two images to be aligned with combined feature areas that overlap each other. The landmark detector generates the points where the two subjects' facial features coincide. Because it is very unlikely that the geometry of this process matches, existing morph techniques incorporate an intermediate warping stage [201]. The last stage is to merge the image(s) into a single one. To this end, a linear map over the entire image is applied. To control the degree to which each sample contributes the morph the additivity of the linear map is controlled by a scalar, which is referred to as the blending factor as it determines the amount of contribution made in the morph creation. Small values reflect no morph is applied, for larger values it will be the impostor(s)' face that contributes to most of the image.

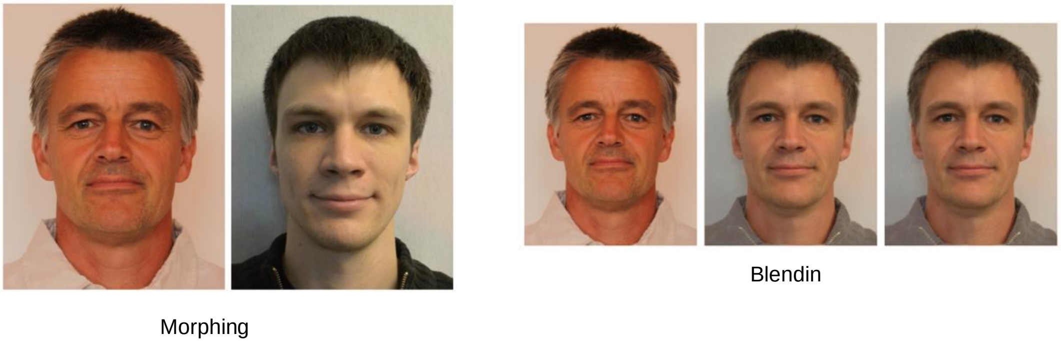


Fig 9: *Morphs attacks, the subjects to be combined (left) and the process using different blending factors.*

Figure 9 shows examples of the morph attack. The associated threat generally implies impersonation. One of the characteristics of the attack vector is that all the subjects involved reveal their identity to some extent. Recently we have observed a trend to use Generative Adversarial Network (GAN) models to produce morphs. These models' foundation rests on the creation of a network that learns the reverse mapping of the generator network present in the gameplay. The method requires three networks, the generator, the discriminator and the translator. Because the generator network produces samples in a noisy feature space, it is necessary to learn what features produce the images that fool the discriminator network, which is the role of the translator network. This network will eventually learn the inversion, using input images to generate features in the latent space for the generator to fool the discriminator. Thus, the three networks are trained simultaneously for the gameplay [4,23]. One of the most common architectures to create GAN morphs is early fusion in the latent space [33]. This technique produces morphs that resemble a target subject from a candidate list by selecting the one that best matches the impostor. Because the generator is trained with the fusions it will output the sample that is most likely to fool the discriminator. This process combines the facial characteristics of both subjects to achieve the win in the GAN gameplay. Recent GAN-based methods allow the creation of samples via controlled migration in the latent space [1]. Thus, it is possible to control the face featuring level in the morph. Fig.10 shows examples of the tuning where it is possible to traverse from gender, ethnicity, age, and hairiness.

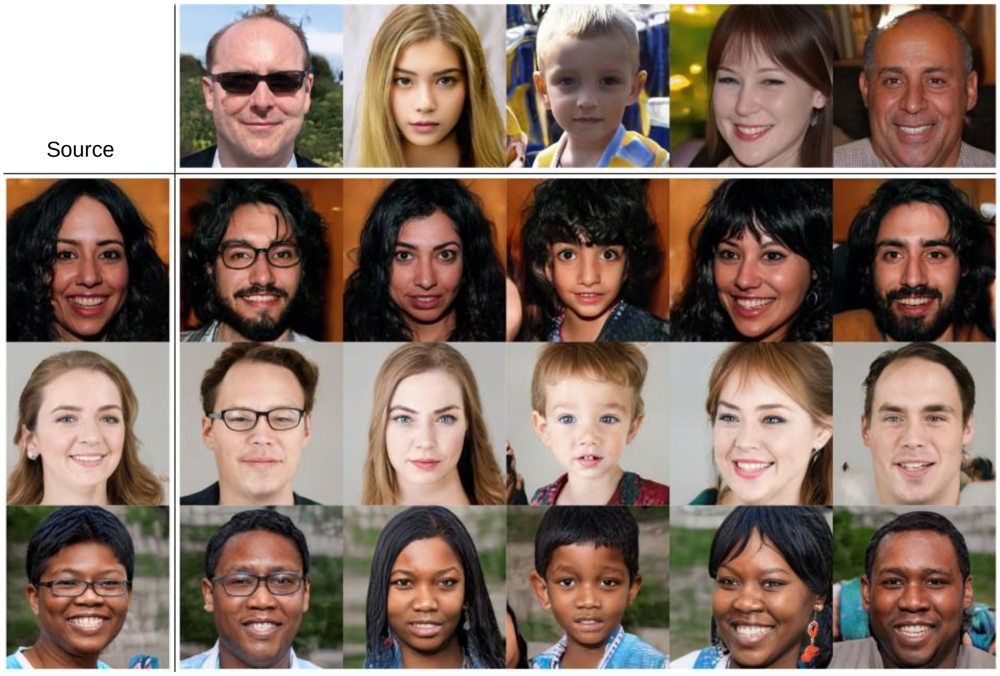
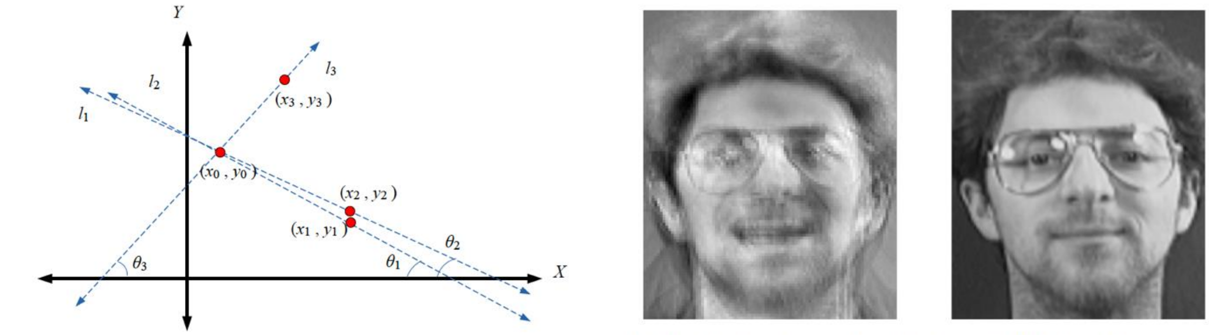


Fig 10: *Morph GAN attacks, the subjects' sources to be combined (first row and first column) and the produced morphs.*

4.8 Template

Template threats exploit vulnerabilities from image features which implies potential vulnerabilities in the feature space. Kaur *et al.* [106] describe how potential attacks, e.g., cross-matching and correlation, may occur in the feature space using simple affine transformations. In this case, features may have very low distances when they are scaled or multiplied by certain vectors. One potential threat is finding samples that closely resemble bonafide templates. In this case, in some feature spaces two nearly identical samples may appear, causing a vulnerability during the matching process [106]. Fig.11a shows an example. Instead of using the original biometric feature vectors, the line's slope and intercepts linking the feature points with random points are used for the matching process. Tampered features may also be produced if the authentication model is known by the attacker [111], while it is even possible to partially recover information that came from the original dataset [175]. Fig.11b shows an example of such an attack, known as an inversion attack. In such a situation, the attacker may reverse the matching process to find templates that cause the system to accept samples via low similarity scores. Template threats mainly aim to bypass the authentication process. It is also possible to discover original samples from the training set by reversing the feature template [113]. Statistical information over a parametric analysis could lead to recovering the samples provided during the training stages [111,175]. It is possible to estimate the input training in a reconstruction process, known as a membership inference attack [111]. Features are also subject to vulnerabilities regarding storage as it is possible to use them in a pre-processed fashion [108]. In this situation, replacement takes place and the input sample has nothing to do with the stored template.



(a) Random Slope to conceal the feature matching process [106].

(b) An image recovered using a new model inversion attack [175]

Fig 11: Template matching threats.

4.9 Model

Model threats regard potential attacks aimed at features processing. Model reliability is essential for high Receiver Operating Characteristics (ROC) performing systems supporting high-security environments. Model threats have shown a recent exponential trend [202], whereas recent work deemed them useful as an Indicator of Compromise (IOC) due to their data nature [173]. Models can be targeted in many ways, and we observe two major categories to create true positives: corrupting the data and creating adversarial samples. Corrupting the data allows the attacker to pass the system authentication using pre-identified authentic samples. Creating adversarial samples requires the attacker to fabricate samples to fool the model, even basic techniques including adding slight perturbations and biometric injection have proven to be successful [203]. We can separate models' attacks into five categories.

4.9.1 Poisoning

Off-the-shelf datasets pose a severe threat as it is possible to insert corrupted data, e.g., people wearing glasses, into the dataset. These eventually create poisoned models that allow certain features to trigger on the input samples [114,125] and match them as authentic. Data can be poisoned in many ways. One common strategy is to mix the target sample with fixed patterns (e.g. another image or random noise) controlling the composite by a factor. This strategy is known as a blended biometric injection attack, Fig.12 (left) shows an example using the hello kitty cat composite. Other strategies, known as physical key attacks [114], require placing accessories over the samples to use them as triggers. These strategies are meant to confuse the FRS into creating incorrect identification. Fig.12 (right) shows examples of the physical keys attack, where the FRS incorrectly matches a subject wearing black glasses.

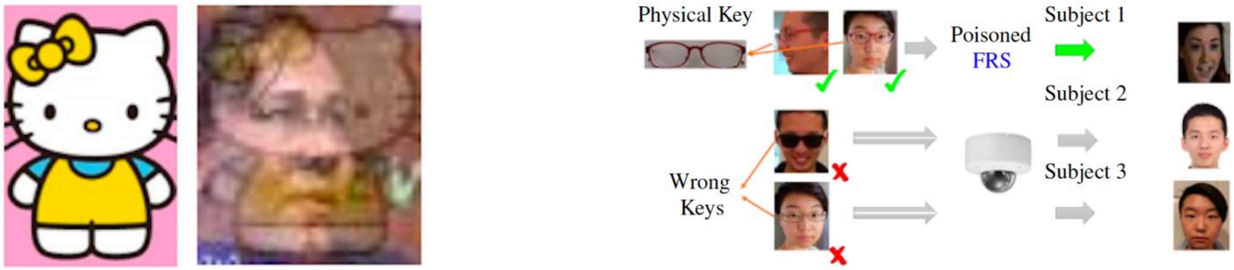


Fig 12: Poison attacks, (left) Blended biometric injection attack (right) physical keys attack.

4.9.2 Backdoor

Model threats are significant in terms of gaining access to the system via non-intrusive backdoor attacks that exploit the fact that fine-tuned models are weakly protected when the FRS relies on third parties' data resources [114,204]. A backdoor attack tricks the model to associate a backdoor pattern with a specific target label, so that, whenever this pattern appears, the model predicts the target label, otherwise, behaves normally. There exist two types of backdoor attacks: 1) tampered-label attack which modifies the label to the target class, and 2) clean-label attack which does not change the label [136]. The threat is severe as even one pixel may be enough to mount the attack [145,147]. The attack requires setting patterns in the form of patches specifically located in the images for the training set and creating a model. The attacker then produces a new image comprising the patch to bypass the FRS. This attack is possible whenever the FRS model is outsourced. Fig.13 shows the attack’s flow, where the trigger allows the attacker to wrongly classify one subject.

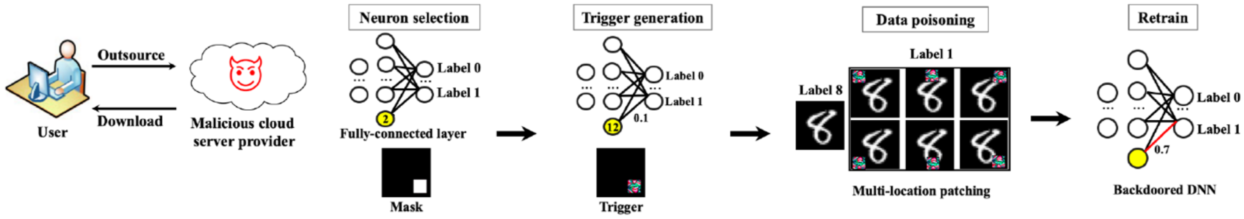


Fig 13: *Backdoor attack. The corrupted model with tampered samples is trained and downloaded from an untrusted party, opening the door to impersonation attacks.*

It is also possible to target specific parts of the model, using strategies known as targeted weight perturbations [137], where the attacker fine-tunes the model for impersonation or concealing purposes. Another possible threat is to produce spurious labels and introduce them to the training set. In this case, one subject may have been assigned more than one label. The FRS will mistakenly classify the subject as another [132], effectively creating a backdoor attack.

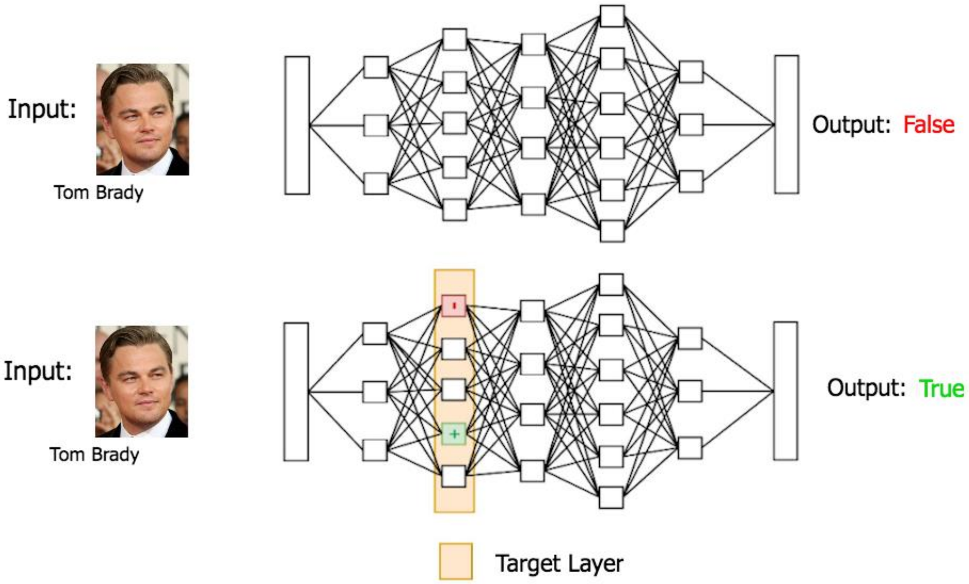


Fig 14: *Weight Perturbation attack. The corrupted model has tampered weights for impersonation/concealing purposes.*

4.9.3 Evasion

These techniques solve a minimisation problem to find the minimum values needed to add noise to the dataset to cause classification errors. Whenever the attacker has plenty of information about the model used by the FRS, it is possible to create a bypass via small perturbations in the presentation samples [117]. To this end, it is necessary to solve an optimisation problem where the minimum perturbation is needed to misclassify the input sample. These strategies are known as evasion attacks [117,203]. Fig.15 shows different purposes of the evasion attack. In a Targeted attack, the adversary aims to change the output from a predefined target label. In a Non-targeted attack, the attacker's goal is to change the results to the labels that are different from the reference data labels, no matter what the label is. In Visual Similarity attacks, the adversarial example resembles input patterns, while in a Non-visual Similarity attack the adversarial example is visually different from the inputs. In this case, the adversarial example is usually noisy or it is a combination of other inputs. Grouping the evasion attacks by their methodology as [205,206] produces:

- **Optimisation:** The goal of the optimiser is to find the minimum adversarial perturbation to the input image samples which causes the model to classify it as belonging to the target class. The perturbation is determined by line-search.
- **Fast Gradient:** The perturbation is found by following the direction of the gradient. The adversarial examples are not computed iteratively but, in a one-step, gradient update along the direction of the gradient sign at each pixel. Commonly, these methods are categorised as one-step methods.
- **Greedy:** Finding the perturbation requires clipping the values of the pixels within range. The method iteratively adjusts the direction that increases the loss of the classifier by running multiple small steps. At each iteration, the values of the pixels of the image are clipped.
- **Region-based:** The assumption is that the input image is located in a region confined by an affine classifier's decision boundaries. At each iteration, the image is perturbed by a small vector. It is sought to lead the resulting perturbed image to the boundaries obtained by linearly approximating the region's boundaries within which the image resides. The perturbations are added to the image and accumulated to compute the ultimate perturbation. This alters the input image label according to the image region's original decision boundaries.
- **Jacobian:** Contrary to perturbing the whole image, the perturbations are with few pixels in the image that might induce significant changes to the output. A saliency map monitors the effect of changing each pixel of the clean image on the resulting classification. The proposed algorithm is repeated until the maximum number of allowable pixels are altered in the adversarial image so that the neural network fooling succeeds.
- **One pixel:** This technique requires the probabilistic labels predicted by the targeted model marginalising any information about the network parameter values or gradients. It is implemented in a simple evolutionary strategy yet successfully fooling networks.

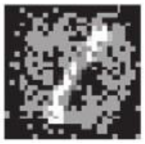
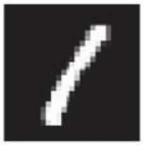
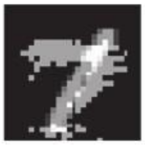
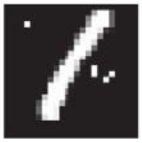
	Non-targeted		Targeted	
Original Image	Non-visual Similar	Visual Similar	Non-visual Similar	Visual Similar
				

Fig 15: *Evasion attack illustration samples. Depending on the purpose the attacker may spear the attack to one particular subject and create visual similarity with the target victim. However, similarity is not always a priority for the attack vector.*

4.9.4 Inversion

Publicly disclosing the model could lead to a threat known as model inversion [172]. This strategy consists of estimating the training dataset(s) via reversing the latent space. If successful, the identities are revealed, and the attacker can produce presentation attacks later [207]. The attack vector utilises the FRS' deployment outputs to gather some basic information regarding their models. This basic information is fed into a reverse analysis followed by the leakage of privacy data embedded in target models, e.g., the reference data (detailed explanation below). Depending on the degree of knowledge of the implemented models, this type of attack can be generally classified into two groups, namely white-box and black- box attacks [203]. In the white-box attack, the attacker can freely access and download the models or other supporting information, while the second one refers to the fact that the attacker only knows the deployment opened by learning models and some information after feeding inputs. This type of attack introduces severe impacts on data privacy.

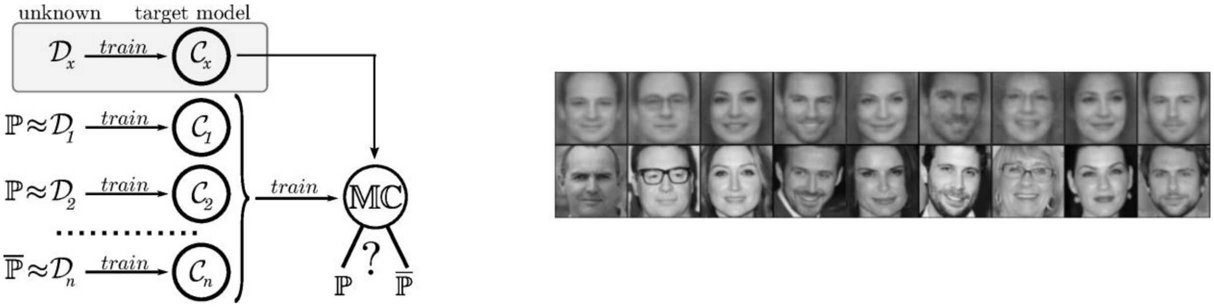


Fig 16: *(left) Property inversion, the attacker infers structural data from the training set. (right) The attacker infers the reference data.*

Fig.16 (left) shows the property inversion attack. The trained classifier C_x tries to infer some statistical information about the training set D_x from which P is the property that the attacker wants to learn about. P could be, e.g., the entries of the training set are equally balanced between males and females. The attacker creates meta-classifier MC trained over the dataset D = D_1, D_2, . . . , D_n with balanced amount of instances reflecting P. The attacker uses the meta-classifier MC to predict which class belongs to the classifier C_x. This is already a new form of information leakage since the adversary learns whether the original training data D_x preserves P

or not. In practice, the attacker is able to infer any key statistical property *P* preserved by the training set performing a sort of brute-force attack on the set of properties. It is important to remark that with this methodology the adversary extracts external information, not in the form of attributes of the dataset D_x. These are essentially statistical properties inferred from the relationship among dataset entries. However, it is possible to recover partial information from the reference data directly from the predictions.

In a reverse model process, the input of the adversarial model is the output of the original model along with its latent space representation. The attacker then minimises the loss between the reconstruction loss of the restored image and the original image while parameterising the observed output (label or class). This process allows associating the predicted class and the input that generated it in the first place. Fig.16 (right) shows reconstruction samples from this process. Naturally, the attack vector requires the baseline of the FRS models.

4.9.5 Registration

It is important to consider the samples that might eventually cause the system to be more vulnerable to adversarial attacks. Before the dataset generation, samples can be classified as either stable or unstable via statistical analysis to avert potential attacks. [170]. The dataset's characteristics play an essential role in preventing spoofing attacks. We find two categories for registration attacks. The first (type 1) is motivated by the idea that an adversarial actor may want to make a small change to the reference data causing a stable input to become unstable. Essentially, adversarial actors may gain access to design files and make undetectable changes. Small changes may lead to catastrophic events such as unstable crack propagation or buckling, whereby an adversarial actor will sample the direct model (ground truth) a fixed number of times, and extract the information needed to create an attack metamodel by training a machine learning model on the data samples. The attack metamodel will identify stable designs. Then, for each stable design, it will test other designs where each altered design is identical to the original one except for one entry. A successful attack predicts that a sample will go from stable to unstable by changing a single entry and is correct with respect to the ground truth. Fig.17 (top) summarises the process.

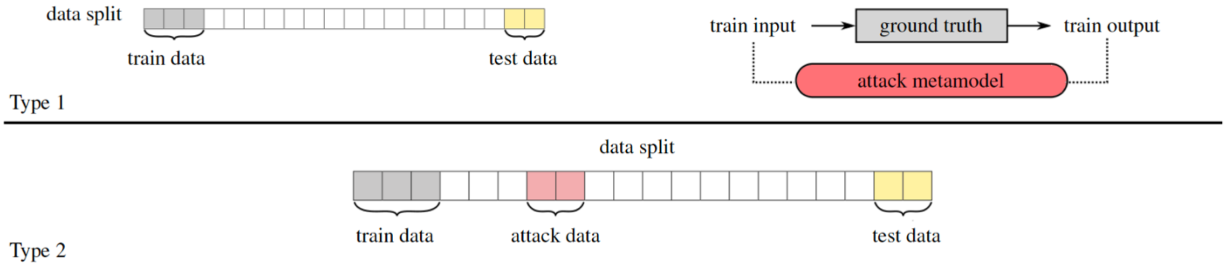


Fig 17: *Registration attack process. The dataset comprises a weak structure for the training/testing stages. (Top) Type 1 attack: a number of model evaluations are required to build an attack metamodel. (Bottom) Type 2 attack: a data subset is run through both the ground truth model and the trained metamodel.*

For the second type, the motivation is that once a metamodel is trained, an adversarial actor may want to identify samples that will deliberately fool the metamodel. Specifically, a type 2 attack is an attack where an adversarial actor has identified input samples that according to the ground truth will be classified as unstable, but a trained metamodel of interest will incorrectly classify it as stable. Given the ground truth and a trained metamodel, a line search over ground truth simulations is used to create a set of classified “attack data”. Then, the ground truth class of the attack data is compared to the predicted class of the trained metamodel. Based on this comparison, it is possible to identify the input samples where the ground truth class is unstable but the trained metamodel prediction is stable. Finally, an attack model is trained to predict these samples. Fig.17 (bottom) illustrates the target data. Few samples, image scale, and insignificant variance are factors that propel vulnerabilities [171], all these factors cause the model to have poor authentication capacities which opens the door to adversarial attacks. Thus, the data must be as inclusive and vast as possible. In general, model threats are feasible because of the ever-growing need for pre-processed data. The lack of control of the training process, reference data and model structure are fertile ground for model vulnerabilities.

4.10 Hardware Attack

Although less frequent, hardware threats pose a serious risk when we observe a trend in distributed computing for mobile devices. Security flaws at the automated electronic design can lead to potential backdoor and hardware trojans [181], which can eventually override the image capturing process. Hardware threats also comprise the perturbation of imaging sensors due to the growth of manufacturers. Unseen sensors may cause system bypasses, which poses a problem for well-established presentation attack defences [179]. Hardware threats are closely related to intended malfunctions in electronic components. An intentional failure may cause the system to produce an undesired output. However, in the capture device case, it is possible to completely avoid this process and send fabricated/stored samples via backdoors [181]. Hardware is particularly weak against side-channel attacks [177], where it may expose the device’s access leading to eventual overrides. Such a situation leaves open doors to take control of the device.

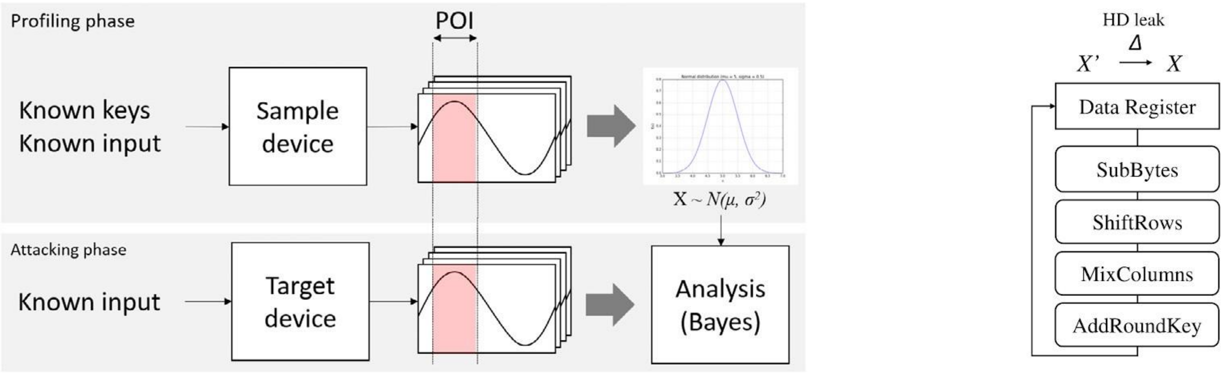


Fig 18: Hardware attack. A side-channel attack to reveal the cipher method used to encrypt the data by the device. (left) Bayes analysis to determine the power consumption probabilistic model. (right) AES transition architecture, where the operations’ loop combined with the Bayes model is used to unveil the cipher.

One side-channel strategy is when the attacker measures the power consumption and determines its probability density function (PDF). This stage is known as the profiling phase, where the attacker operates a target cryptographic module and captures a power waveform and calculates the conditional probability of each estimated key with the PDF built in the profiling phase. Later, the attacker applies the probabilities and unveils the secret key (attacker phase). When given templates correspond to each estimated key, the attacker can calculate the probability of the observed waveform. This refers to the probability that the estimated key is used under the condition where the observed waveform is happening. Based on this approach two common power consumption leak models are as follows: (1) the dominant factor of the power consumption in a CMOS device is its switching activities. The Hamming distance of the leak model's objective is to count the number 0/1 transitions of the bit sequence of an intermediate value involved with the secret key. When an encryption circuit has a loop architecture such as AES, the power consumption is considered proportional to the number of bit transitions in the data register that holds an intermediate value (2). The Hamming weight leak model has a slightly different methodology to count the number of ones in a bit sequence being processed. Fig.18 shows the typical AES algorithm loop architecture, where the data registers hold the intermediate value. Another potential threat is the recent growth of electronic manufacturers. Having the foundation in digital forensics using Sensor Pattern Noise [208], electronic devices produce unique noisy patterns. Those patterns then can be used to produce samples from non-trusted sources, imitating samples taken from a different electronic device. The imaging sensor may cause systems bypass in this fashion [179] when the attacker discovers the victim's device model. Hence the importance of building datasets comprising as many electronic devices as possible. This group represents a growing threat in mobile devices as new technology trends migrate part of the FRS processing to remote devices [209–211]. This migration delegates these devices as important security points.

4.11 Transmission/Storage

The data transmission and storage group regards the potential threats during the data handling process. Common threats include data replay and data breaches via man-in-the-middle attacks [187]. Intuitively these threats are not exclusive to FRS data, but common to data encryption protocols. Harkeerat-Khanna [182] observe that biometric features require protection from the mathematical perspective. One of the most severe threats in this group is the potential consequences for compromised databases. If not properly mitigated, addressing it may require a full user re-enrolment. Fig. 19 shows vulnerable points of generic remote authentication systems. Remote authentication systems are at greater security risk as they are difficult to supervise and practice control over the claimant and the verifier. In a server- spoofing attack, the attacker manipulates the claimant by communicating as a valid verifier using an IP address or DNS server fraudulently to acquire victims' biometric information. These are the most common means of online identity theft. In eavesdropping, the attacker hears the transmitted signal, while in a replay attack, the attacker records and resubmits the signal to gain access. In attacks-via-record multiplicity, the attacker obtains multiple different templates belonging to the same user eventually revealing identities. In man-in-the-middle attacks, it is also possible for an adversary to insert himself into the communication link and impersonate both parties to gain the transmitted information. The matching process and result can be corrupted and overridden by the use of malicious software, e.g., Trojan horses. The database can be hacked and analysed for personal information and cross-matching attacks. Further, privacy is threatened as the individual's control over the collection, use, and disclosure of his biometric data is compromised.

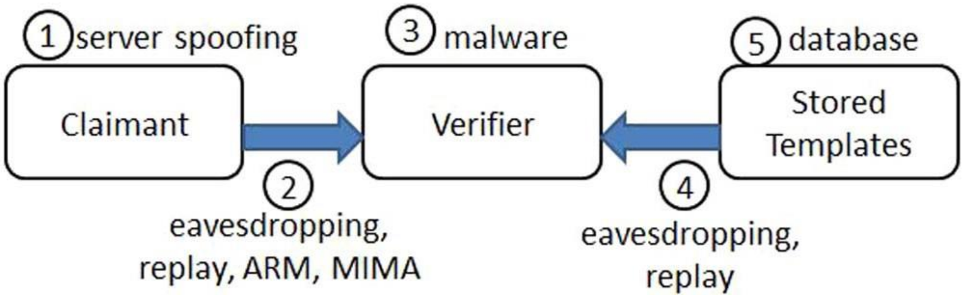


Fig 19: *Transmission/Storage attacks. A generic biometric system can be targeted during the transmission and storage operation compromising the entire system.*

Finally, database cross-matching allows determination of the relation among reference templates for unintended activities, such as covert surveillance or profiling. The unintended use of biometric information for applications other than authentication also puts user privacy at serious risk. Transmission/Storage threats are present during information processing. Open insecure channels facilitate replay attacks while impersonation is possible if the data is not properly tokenized [182,186]. This vulnerability is present for all the data involved: models, raw images, parameters, etc. For communications tools, FRS threats can exploit vulnerabilities in the middle layers of the OSI model. Man-in-the-middle attacks can potentially replace the end-user [187,190] to establish communication with the impersonator. The most severe threat in this group is backdoor access. If the attacker manages to gain administrator-level credentials, the whole system is compromised [194] as the attacker has direct access to every resource, and entire destruction of the FRS is possible.

5 PRISMA Analysis

The number of papers referring to each category were counted to elaborate a chart. Fig 20 shows the attack proportions. In this figure, we observe the most persistent threats are for the presentation stage.

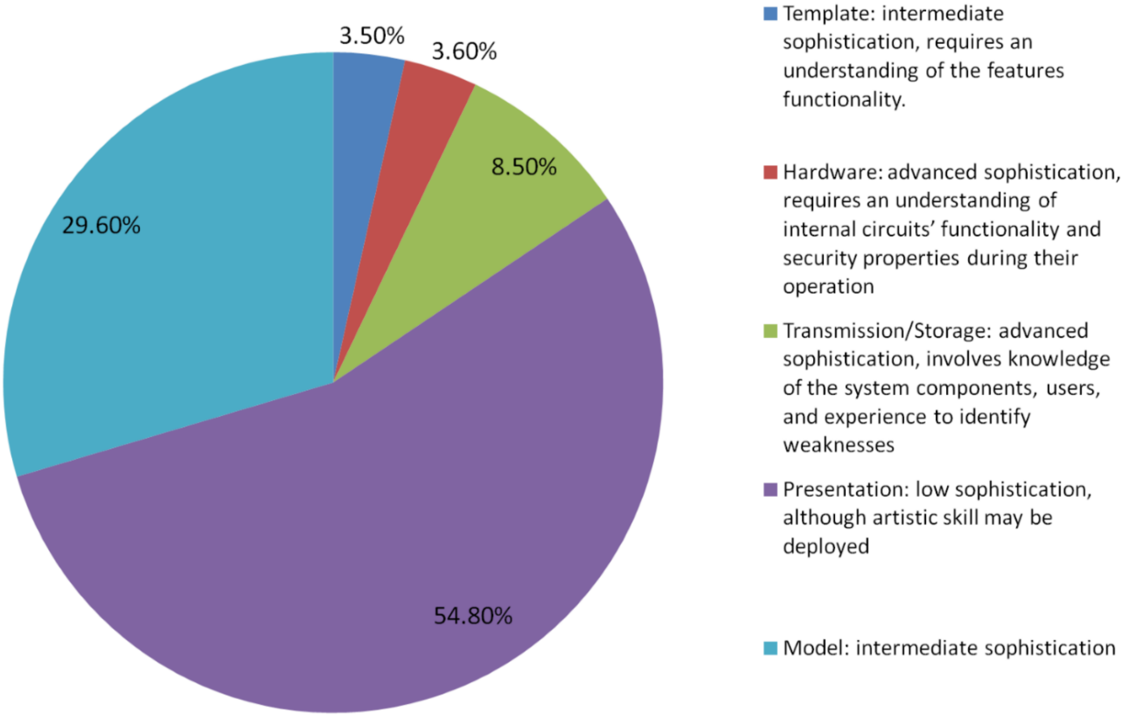


Fig 20: *Attacks by proportion groups. From 2018–2021, studies focus on the presentation attack more frequently than any other sort of attack*

It is noted that FRS are mostly targeted during the presentation stage. The methods are highly related to their level of sophistication, and thus, to the system stage in which they take place. Despite highly sophisticated attacks that may appear at any stage, the more elaborated attacks persist as those attack vectors that gain access to internal components of the system. This is particularly the case with Transmission/Storage attacks where the attacker gains control of every component or its functionality. The following list details the level of sophistication required:

1. Presentation: low sophistication, the attack vector may require only a high-quality printer, masks, or electronic device output. However, some require extraordinary artistic skills.
2. Template: intermediate sophistication, the attack vectors require an understanding of the FRS features functionality.
3. Model: intermediate sophistication, some attack vectors require the victim to download a tampered model.
4. Hardware: advanced sophistication, attack vectors require an understanding of internal

circuits' functionality and security properties during their operation.

5. Transmission/Storage: advanced sophistication, the attack vector necessarily involves knowledge of the FRS components, users, and experience identifying weakly protected systems and networks.

Table 2 summarizes the PRISMA literature review.

Table 2
The table shows the main details related to each category.

Sophistication	Access Level	Severity	Attack vector	Devices/Resources
Low	External	Low	Presentation	Printers, electronic devices, and raw materials
Intermediate	External/Internal	Low	Template	Computer
Intermediate	External/Internal	Medium	Model	Computer
Advanced	External	Medium	Hardware	Mobile device, Computer
Advanced	Internal	High	Transmission/Storage	Computer

The table comprises the attack vector, which is the means an attacker follows to achieve a goal. The level of sophistication of the attacker reflects the technical background required to perform the attack. The access level reflects where the means of the attack needs to take place. The severity is the harm the attack can cause the FRS. The devices or resources required to perform the attack are also listed.

6 Authentication Reference Architecture

The Authentication Reference Architecture (ARA) refers to the structure and interconnectivity of the FRS. It comprises the key components and their usage in the design, allowing the visibility to identify the possible attack points. In this section we present reference architectures commonly used to detect vulnerabilities.

6.1 IBM

These threat groups are present in several threat models. In 2001, IBM initially proposed a generic biometrics-based system model for analysing security and privacy in authentication systems [212]. This proposed model, presented in Fig 21 consists of four key components, sensor, feature extraction, matcher and store template(s), and eight potential risk points are associated with these components. Some possible attacks which may occur in these points are:

1. Presenting fake biometrics at the sensor.
2. Resubmitting previously stored digitised biometrics signals.
3. Overriding the feature extraction process.
4. Tampering with the biometric feature representation.
5. Corruption of the matcher.
6. Tampering with stored templates.

7. Attacking the channel between the stored templates and the matcher.
8. Overriding the final decision

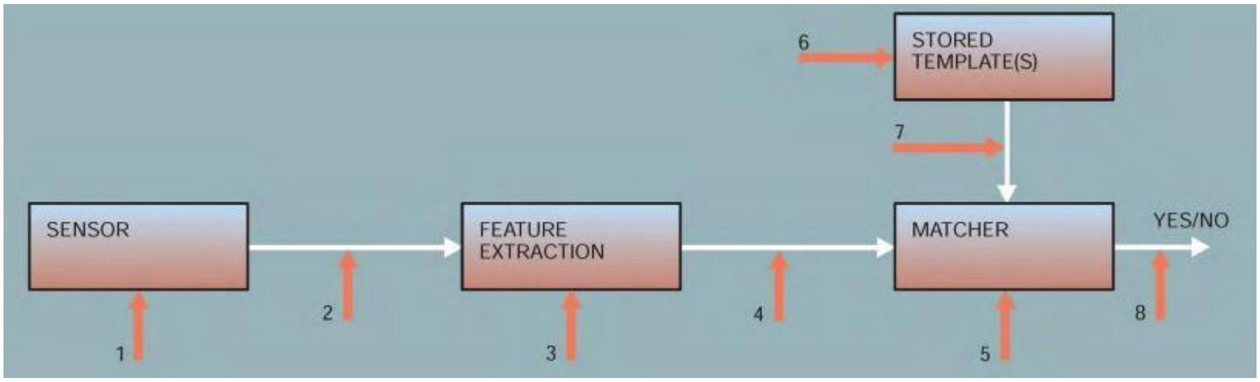


Fig 21: *Possible attack points in a generic biometrics-based system as proposed by IBM*

6.2 NIST

Although the model proposed by IBM covers most components related to a biometrics-based system, an increasing number of features and processes in such systems are found with potential risks. In 2015, the National Institute of Standards and Technology (NIST) defined a refined system architecture to model biometric system attacks [213]. This architecture shown in Fig.22, includes five critical components and eleven risk points. Compared with IBM's model, two decision-making related and one data-capture related risk points supplement this model, reflecting the recent increase in the popularity of biometric technologies on mobile devices, which work in unsupervised conditions.

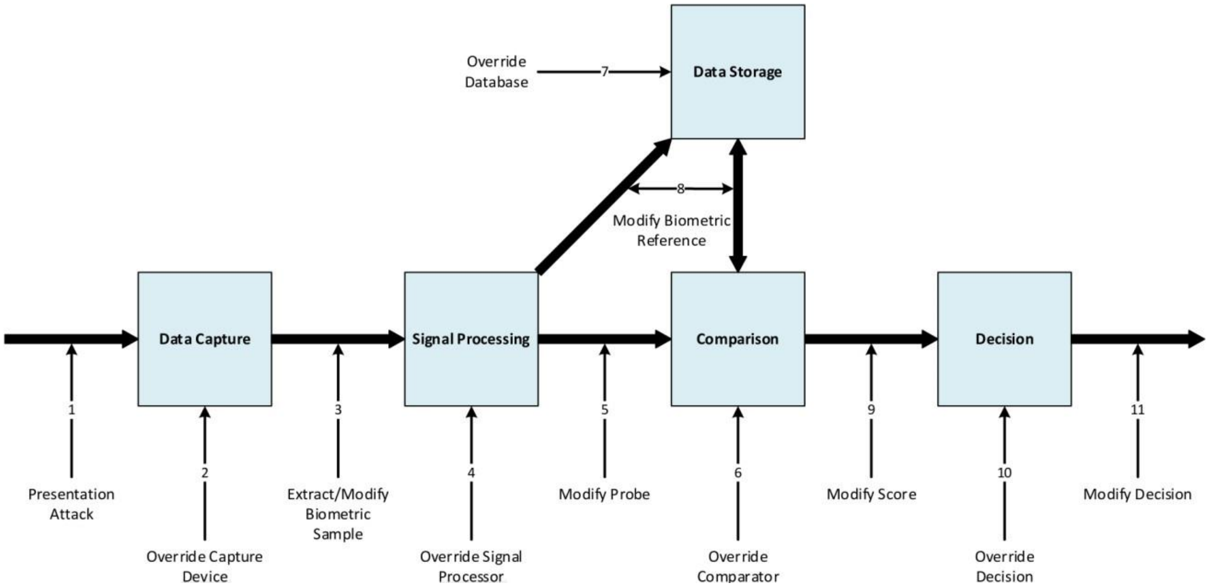


Fig 22: *The Authentication Reference Architecture as proposed by NIST .*

1. Presentation attack. The face images are not secret, and they may be available on social media and company websites. The spoofing attacks are challenging in unsupervised environments, such as mobile devices and distributed authentication systems, where no operator monitors the biometric capture.

2. Override capture device. Due to the increasing number of biometric technologies applied in mobile devices, attackers may modify or override the capture devices. The devices send the modified rather than the captured image to the following step.

3. Extract/Modify biometric sample. The captured samples are transmitted for signal processing or quality assessment. The transmission bus should be secure if biometric capture and signal processing are executed in the same device. Moreover, if the biometric capture device is mobile and the captured data is transmitted to the server for further processing, man-in-the-middle attacks may be performed.

4. Override signal processor. Signal processing often follows capture. The signal processor, usually an algorithm or application, extracts essential information from the captured data and then constructs a template. Attackers may replace the feature extraction algorithm with a new one, generating a fake feature for a special user to pass the system without authentication.

5. Modify probe. Like the modification of biometric samples, attackers may reverse the features/templates to reject the authentication that should be accepted. The potential risk in this step is always associated with the attack approaches in other steps.
6. Override comparator. The comparator, the algorithm or application used in generating match scores between sets of biometric samples is essential to a successful biometric authenticator. The attacks and risks in this point may include an inaccurate algorithm or the algorithm with Trojan or backdoor.
7. Override database. Data storage is the crucial component of the system. The raw images and the generated templates are stored in a centralised database within a distributed database. An attacker may modify/add/remove some raw image data or compromise the generated ones.
8. Modify biometric reference. Similar to modifying the biometric sample, the biometric reference can also be modified during transmission.
9. Modify score. Attackers may change the score of comparison before it is sent to the decision engine.
10. Override decision engine. The final decision is made by comparing the score and the pre-defined threshold.
11. Attackers may modify the threshold to pass authentication and put the system at a low-security level.
12. Modify decision. The easiest way to pass the identification system is to modify the final decision directly. It is crucial to protect the transmission of the decision to the requesting entity, such as devices, systems or applications.

6.3 ISOTEC

ISOTEC [214] elaborated an FRS architecture for biometrics in 2016: ISO/IEC 30107-1:2016. Identifying nine potential vulnerabilities as Fig.23 depicts.

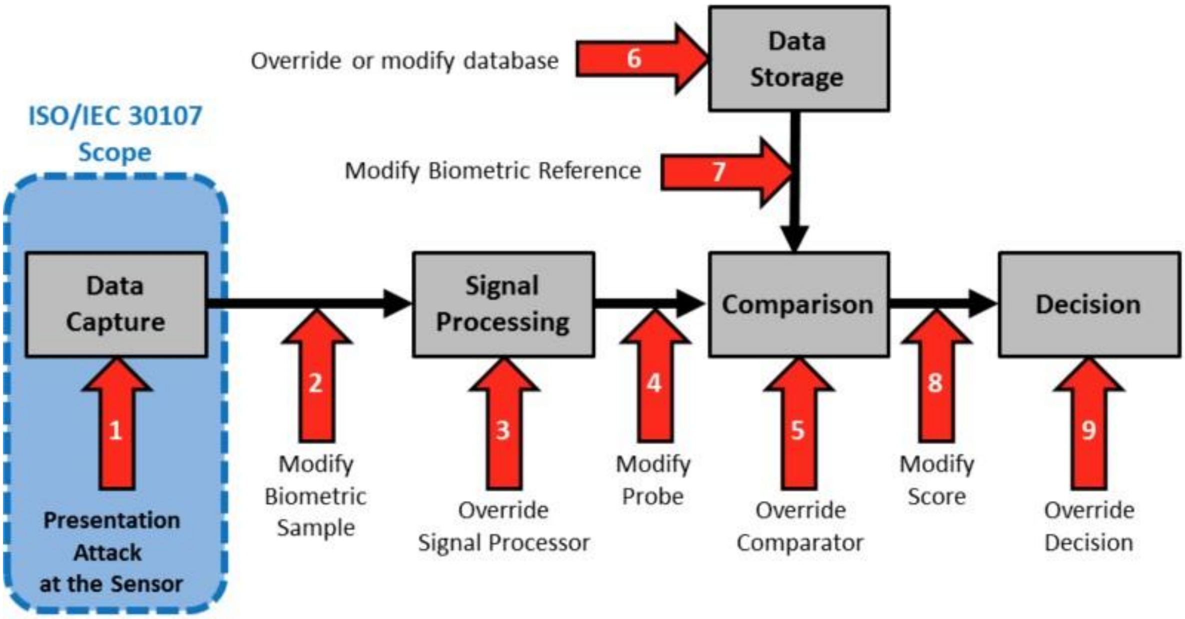


Fig 23: *Vulnerability of a face recognition system ISOTEC 2016.*

1. Vulnerability is noted at the sensor and involves presenting a face biometric artefact of the legitimate user as an input to the sensor. An artefact is defined as an artificial object or representation presenting a copy of biometric characteristics or synthetic biometric patterns. This kind of attack is known as a presentation attack and is defined as a presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system.
2. Vulnerability related to intercepting the biometric sample that was captured by the sensor. This attack involves replacing the captured face biometric sample with a fake sample.
3. Vulnerability overrides the signal processing module. This could involve modifying the functionality of the feature extractor, for instance, using a Trojan Horse.
4. Vulnerability allows the attacker to replace the extracted features of the probe sample with target features.
5. Vulnerability involves overriding the comparator so that it will output a comparison score required by the attacker.
6. Vulnerability involves replacing the reference template such that the authorized ID is associated with the attacker template.
7. Vulnerability is the modification of the reference template in the communication channel.
8. Vulnerability is the interception and corruption of the comparator output.
9. Vulnerability involves overriding the decision module to output the intended decision.

Of these nine vulnerabilities, only the first involves an attack on the sensor itself; all the other vulnerabilities are related to the integrity of the overall system. Attacks on facial sensors have garnered broad interest from the biometric community, as using this approach: it is easy to attack a biometric system, it is easy to generate the face artefact and to present it to the sensor, and it does not require knowledge about the operational details of the biometric system.

7 Threat Actors

Threat actors are those individual(s) and organisation(s) that actively pose a threat to specific resource(s). These have malicious purposes by the means of taking advantage of security vulnerabilities, inappropriate security awareness, or fragile technological deployments to corrupt the systems' functionality [215]. The motivations are multiple and vary, e.g., fraud, geopolitical gains, affecting the users' data, or even personal satisfaction [216]. Cyber threat actors are not equal in terms of capability and sophistication and have a range of resources, training, and support for their activities. They may operate on their own or be hired as part of a larger organisation. This could be the case of nation-state intelligence spy bureaus or an organised crime group. Some of the actors targeting FRS and their motivations are described below.

7.1 Cyber-criminals

Most cybercriminals are financially motivated, where the ultimate goal is to generate revenues from their activities. Cybercrime has been growing recently because of the COVID-19 pandemic as a significant workforce migrated to work remotely. In isolation, cyber-criminals are generally deemed to have moderate technical sophistication. Nonetheless, when organised they plan and support attack vectors in addition to specialised technical capabilities that affect a large number of victims. They create sophisticated structures to offer services and gain knowledge to improve their skills via forums and encrypted-protected websites, and provide Cybercrime as a Service(CaaS), where criminals without technical background share the profits, in exchange for tools that perpetrate attacks [217]. This combination has led to an exponential trend of crime, and a particular trend in Ransomware as a Service(RaaS), where criminals are enabled to hijack and hold their victim's data ransom until they pay for an encrypted data key to regain access to it.

7.2 Insiders

Insider threat actors are people working inside the organisation. They can be financially motivated, for instance to commit fraud, or be unhappy or offended employees. Insider threats are particularly dangerous because of their access to internal systems and networks that are protected by physical and virtual security perimeters. The steps needed to gain access to crucial resources and components are given away eliminating the need to employ other remote means [215], e.g., the steps followed by cyber-criminals.

7.3 Thrillers

There are threat actors that are motivated for reputation and satisfaction. These do not need profits for perpetuity. The aims seem to be to become symbols or earn community acknowledgment and admiration. Thrillers are commonly low-level sophistication threat actors as they often rely on widely available tools that require little technical skill to deploy [215].

7.4 Hacktivist

Ideologically motivated actors are known as Hacktivists. They have the purpose of transmitting an idea by the means of disrupting the intended actor, organisation, or states. The ends are mainly promoting a political agenda or social change. Hacktivists also draw attention to issues or expose wrongdoings, privacy violations, eco-crimes, censorship, and illegal activities from malicious actors with the aim of punishing the relevant authorities. The most common Hacktivists' practice is to publicly disclose information from their victims, thereby creating a severe impact facilitated by those using the information for secondary purposes, e.g., journalist, cyber-criminals, nation states. Hacktivists in general create sophisticated groups as their activities require organisation and the ability to gain access to well-protected information.

7.5 Terrorist

Similar to Hacktivists, terrorists are ideologically motivated. However, they tend to deploy violence, using cyber-crime to fund their activities. It is being reported that very often terrorists use freely available tools which require little or no technical skills to perform attacks. Thus terrorists are deemed low level of sophistication threat actors [216].

7.6 Competitors

Companies compete in many ways; one possible way to outperform a competitor company is by disrupting its activities and putting it out of business. As other common practices, such as in support of short-selling cybercrime can be used as a practice to position a company over another. This threat actor can cause severe damage by, e.g., making a competitor website unreachable, disclosing its clients' private information, spreading misinformation, or revealing structural information.

7.7 Nation-States

Advanced Persistent Threats (APT) represent those threat actors in the top tier of sophistication and skills, capable of using advanced techniques to conduct complex and protracted campaigns in the pursuit of their strategic goals [215]. This threat actor regularly involves nation states or very proficient organised cyber-crime groups. Nation-state cyber threat actors are often geopolitically motivated and are the most sophisticated actors. They have dedicated resources and personnel, and extensive planning and coordination [216]. Some nation states have operational relationships with private sector entities and organised criminals. They have the capacity to spear-target politicians, interfere with economies, disrupt governments' critical infrastructure collapse businesses or individuals' economic activities, steal medical advances, and carry out cyberwarfare. Nation-state threat actors also have the capacity to organise ramifications with proxy actors, where cyber-criminals, terrorists, and hacktivists act as mercenaries or subsidisers for elaborated purposes. In such cases, the principal actor is not visible, and it is difficult to hold them accountable for criminal activities for which they ultimately benefit. Table 3 summarizes the threat actors and crucial details relating to their characteristics.

Table 3
Threat Actors. The table shows the characteristics of the threat actors.

Threat Actor	Sophistication	Motivation	Severity
Cyber-criminals	All	Financial	Medium
Insiders	Low	Financial/Revenge	Low
Thrillers	Low	Satisfaction	Low
Hacktivist	Intermediate/High	Ideological	Medium/High
Terrorist	Low	Ideological	Low
Competitors	Low/Intermediate	Financial	High
Nation-States	High	Geo-political	High

8 Conclusions

This work particularly aims to expose the threat taxonomy and categorise the attacks of recently discovered vulnerabilities against Facial Recognition Systems. We have presented an extensive literature review of recently discovered attacks and their place in the Authentication Reference Architecture for these systems. Three key findings of this review are:

We argue that the presentation stage needs to be the most secure part of the system. The presentation stage of the system presents most vulnerabilities in proportion to other stages in the processes as attackers do not need high-level technical skills to cause harm. For instance, a masquerade or deceiving face can be created using makeup. Presentation attacks therefore could cause severe distrust towards FRS.

Some pre-trained models and datasets used to save time and increase accuracy of FRS open the door to several attacks. We understand that some techniques require elaborated mathematical analysis of the models' functionality, e.g., backdoor learning. However, other approaches require only elementary data transformations to mount classification errors, e.g., spurious labelling. An open-source online repository could host corrupted datasets or pre-trained models, which could cause severe damage. Intuitively the level of expertise to create such an attack does not require an elaborated technical background.

Mobile devices are gaining ground In FRS. We observe delegation to mobile devices in many user-identification systems, for example in mobile banking, which requires respective security measures. We identified potential vulnerabilities in mobile devices that in the near future may be increasingly exploited due to this tendency. In particular, we estimate that device overriding concerns may increase both because of the technology and fast-growing adoption rates.

Keeping in mind these findings, next steps for our research is to focus on threat models and authentication reference architectures for face recognition systems to support robust design and trustworthiness of face recognition systems.

9 References

[1]T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[2]U. Scherhag, C. Rathgeb, and C. Busch, “Towards detection of morphed face images in electronic travel doc- uments,” in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, April 2018, pp. 187–192.

[3]H. Bian, D. Chen, K. Zhang, H. Zhou, X. Dong, W. Zhou, W. Zhang, and N. Yu, “Adversarial defense via self-orthogonal randomization super-network,” *Neurocomputing*, vol. 452, pp. 147–158, 2021. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S0925231221006044

[4]N. Damer, V. Boller, Y. Wainakh, F. Boutros, P. Terh örst, A. Braun, and A. Kuijper, “Detecting face morph- ing attacks by analyzing the directed distances of facial landmarks shifts,” in *Pattern Recognition*, T. Brox, A. Bruhn, and M. Fritz, Eds. Cham: Springer International Publishing, 2019, pp. 518–534.

[5]L. Kurnianggoro and K.-H. Jo, “Ensemble of predictions from augmented input as adversarial defense for face verification system,” in *Intelligent Information and Database Systems*, N. T. Nguyen, F. L. Gaol, T.-P. Hong, and B. Trawiński, Eds. Cham: Springer International Publishing, 2019, pp. 658–669.

[6]L. Debiasi, N. Damer, A. M. Saladi é, C. Rathgeb, U. Scherhag, C. Busch, F. Kirchbuchner, and A. Uhl, “On the detection of gan-based face morphs using established morph detectors,” in *Image Analysis and Processing – ICIAP 2019*, E. Ricci, S. Rota Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham: Springer International Publishing, 2019, pp. 345–356.

[7]J. Zhou, C. Liang, and J. Chen, “Manifold projection for adversarial defense on face recognition,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 288–305.

[8]S. Wang, U. M. Kelly, and R. N. Veldhuis, “Gender obfuscation through face morphing,” in *2021 IEEE Inter- national Workshop on Biometrics and Forensics (IWBF)*, May 2021, pp. 1–6.
[9]A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, “Fast geometrically-perturbed adversarial faces,” in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan 2019, pp. 1979–1988.
[10]N. Damer, F. Boutros, A. M. Saladi é, F. Kirchbuchner, and A. Kuijper, “Realistic dreams: Cascaded enhance- ment of gan-generated images with an example in face morphing attacks,” in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sep. 2019, pp. 1–10.
[11]A. I. Rohra and R. K. Kulkarni, “Survey on recent trends in image morphing techniques,” in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Nov 2019, pp. 19–23.
[12]L. Qin, F. Peng, S. Venkatesh, R. Ramachandra, M. Long, and C. Busch, “Low visual distortion and robust morphing attacks based on partial face image manipulation,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 72–88, Jan 2021.
[13]H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “Mipgan—generating strong and high quality morphing attacks using identity prior driven gan,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 365–383, July 2021.
[14]H. H. Nguyen, J. Yamagishi, I. Echizen, and S. Marcel, “Generating master faces for use in performing wolf attacks on face recognition systems,” in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2020, pp. 1–10.
[15]D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, and E. Cabello, “Border control morphing attack detection with a convolutional neural network de-morphing approach,” *IEEE Access*, vol. 8, pp. 92 301–92 313, 2020.
[16]P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi, “Attention aware wavelet-based detection of morphed face images,” in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Aug 2021, pp. 1–8.
[17]S. Venkatesh, K. Raja, R. Ramachandra, and C. Busch, “On the influence of ageing on face morph attacks: Vulnerability and detection,” in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2020, pp. 1–10.
[18]S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwers, R. Veldhuis, and C. Busch, “Morphed face detection based on deep color residual noise,” in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Nov 2019, pp. 1–6.
[19]U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, “On the vulnerability of face recognition systems towards morphed face attacks,” in *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, April 2017, pp. 1–6 [20]S. Venkatesh, R. Ramachandra, and K. Raja, “Face morphing of newborns can be threatening too : Preliminary study on vulnerability and detection,” in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Aug 2021, pp. 1–8.
[21]Z. Blasingame and C. Liu, “Leveraging adversarial learning for the detection of morphing attacks,” in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Aug 2021, pp. 1–8.
[22]A. Makrushin and A. Wolf, “An overview of recent advances in assessing and mitigating the face morphing attack,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1017–1021.
[23]N. Damer, S. Zienert, Y. Wainakh, A. M. Saladi é, F. Kirchbuchner, and A. Kuijper, “A multi-detector solution towards an accurate and generalized detection of face morphing attacks,” in *2019 22th International Conference on Information Fusion (FUSION)*, July 2019, pp. 1–8.
[24]L. Wandzik, G. Kaeding, and R. V. Garcia, “Morphing detection using a general- purpose face recognition system,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1012–1016.
[25]Y. Minakawa, M. Abe, K. Sekine, and Q. Zhao, “Neural network based feature point detection for image mor- phing,” in *2014 IEEE 6th International Conference on Awareness Science and Technology (iCAST)*, Oct 2014, pp. 1–6.
[26]C. Seibold, A. Hilsmann, and P. Eisert, “Reflection analysis for face morphing attack detection,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1022–1026.
[27]P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi, “Detection of morphed face images using discriminative wavelet sub-bands,” in *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, May 2021, pp. 1–6.
[28]U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. J. Veldhuis, L. Spreeuwers, M. Schils, D. Mal- toni, P. Grother, S. Marcel, R. Breithaupt, R. Ramachandra, and C. Busch, “*Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting*,” in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2017, pp. 1–7.

[29]U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch, “Deep face representations for differential morphing attack detection,” *IEEE Transactions on* Information Forensics *and Security*, vol. 15, pp. 3625–3639, 2020.

[30]L. Debiasi, C. Rathgeb, U. Scherhag, A. Uhl, and C. Busch, “Prnu variance analysis for morphed face image detection,” in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Oct 2018, pp. 1–9.

[31]R. Ramachandra, S. Venkatesh, K. Raja, and C. Busch, “Towards making morphing attack detection robust using hybrid scale-space colour texture features,” in *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, Jan 2019, pp. 1–8.

[32]S. Jassim and A. Asaad, “Automatic detection of image morphing by topology-based analysis,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1007–1011.

[33]N. Damer, A. M. Saladi é, A. Braun, and A. Kuijper, “Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network,” in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Oct 2018, pp. 1–10.

[34]Z. Ming, M. Visani, M. M. Luqman, and J.-C. Burie, “A survey on anti-spoofing methods for face recognition with rgb cameras of generic consumer devices,” 2020.

[35]U. Muhammad and A. Hadid, “Face anti-spoofing using hybrid residual learning framework,” in *2019 Interna- tional Conference on Biometrics (ICB)*, June 2019, pp. 1–7.

[36]R. Ramachandra and C. Busch, “Presentation attack detection methods for face recognition systems: A comprehensive survey,” *ACM Comput. Surv.*, vol. 50, no. 1, Mar. 2017. [Online]. Available: https://doi.org/10.1145/3038924

[37]K. Kotwal, Z. Mostaani, and S. Marcel, “Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 1, pp. 15–25, Jan 2020.

[38]C. Rathgeb, P. Drozdowski, D. Fischer, and C. Busch, “Vulnerability assessment and detection of makeup presentation attacks,” in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, April 2020, pp. 1–6.

[39]Y. Liu, J. Stehouwer, and X. Liu, “On disentangling spoof trace for generic face anti-spoofing,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 406–422.

[40]X. Yang, F. Wei, H. Zhang, and J. Zhu, “Design and interpretation of universal adversarial patches in face detection,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: [41]L. Chai, D. Bau, S.-N. Lim, and P. Isola, “What makes fake images detectable? understanding properties that generalize,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 103–120.

[42]L. Gao, Q. Zhang, J. Song, X. Liu, and H. T. Shen, “Patch-wise attack for fooling deep neural network,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 307–322.

[43]Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, “Revisiting pixel-wise supervision for face anti-spoofing,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 285–295, July 2021.

[44]K. Kong, X. Hei, T. Zeng, C. Ling, C. Zhang, B. Song, H. Cao, and M. Peays, “A countermeasure against face-spoofing attacks using an interaction video framework,” in *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*, Oct 2017, pp. 758–763.

[45]B. Geng, C. Lang, J. Xing, S. Feng, and W. Jun, “Mfad: A multi-modality face anti-spoofing dataset,” in *PRI- CAI 2019: Trends in Artificial Intelligence*, A. C. Nayak and A. Sharma, Eds. Cham: Springer International Publishing, 2019, pp. 214–225.

[46]D. T. van der Haar, “Real-time face antispoofing using shearlets,” in *Information Security*, H. Venter, M. Loock,
M. Coetzee, M. Eloff, and J. Eloff, Eds. Cham: Springer International Publishing, 2019, pp. 16–29.

[47]A. Khurshid, S. C. Tamayo, E. Fernandes, M. R. Gadelha, and M. Teofilo, “A robust and real-time face anti-spoofing method based on texture feature analysis,” in *HCI International 2019 – Late Breaking Papers*,
C. Stephanidis, Ed. Cham: Springer International Publishing, 2019, pp. 484–496.

[48]R. Bresan, C. Beluzo, and T. Carvalho, “Exposing presentation attacks by a combination of multi-intrinsic image properties, convolutional networks and transfer learning,” in *Advanced Concepts for Intelligent Vision Systems*,
J. Blanc-Talon, P. Delmas, W. Philips, D. Popescu, and P. Scheunders, Eds. Cham: Springer International Publishing, 2020, pp. 153–165.

[49]A. Nema, “Ameliorated anti-spoofing application for pcs with users’ liveness detection using blink count,” in *2020 International Conference on Computational Performance Evaluation (ComPE)*, July 2020, pp. 311–315. [50]S. K. Adari, W. Garcia, and K. Butler, “Adversarial video captioning,” in *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, June 2019, pp. 24–27.

[51]X. Zhao, Y. Lin, and J. Heikkil ä, “Dynamic texture recognition using volume local binary count patterns with an application to 2d face spoofing detection,” *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 552–566, March 2018.

[52]D. Kasat, S. Jain, and V. Thakare, “Real time face morphing,” in *2015 International Conference on Computa- tional Intelligence and Networks*, Jan 2015, pp. 160–165.

[53]I. Buciu and S. Goldenberg, “Oscillating patterns based face antispoofing approach against video replay,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 02, May 2015, pp. 1–6.

[54]D. Timoshenko, K. Simonchik, V. Shutov, P. Zhelezneva, and V. Grishkin, “Large crowdcollected facial anti- spoofing dataset,” in *2019 Computer Science and Information Technologies (CSIT)*, Sep. 2019, pp. 123–126.

[55]A. Agarwal, A. Sehwag, R. Singh, and M. Vatsa, “Deceiving face presentation attack detection via image transforms,” in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, Sep. 2019, pp. 373–382.

[56]M. Ouyang, R. K. Das, J. Yang, and H. Li, “Capsule network based end-to-end system for detection of replay attacks,” in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Jan 2021, pp. 1–5.

[57]J. Zhou, C. Ge, J. Yang, H. Yao, X. Qiao, and P. Deng, “Research and application of face anti-spoofing based on depth camera,” in *2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI)*, Sep. 2019, pp. 225–229.

[58]A. F. Ebihara, K. Sakurai, and H. Imaoka, “Specular- and diffuse-reflection-based face spoofing detection for mobile devices,” in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[59]X. Qu, J. Dong, and S. Niu, “shallowcnn-le: A shallow cnn with laplacian embedding for face anti-spoofing,” in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, May 2019, pp. 1–8.

[60]Z. Yu, X. Li, P. Wang, and G. Zhao, “Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection,” *IEEE Signal Processing Letters*, vol. 28, pp. 1290–1294, 2021.

[61]S. R. Arashloo, “Unseen face presentation attack detection using sparse multiple kernel fisher null-space,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4084–4095, Oct 2021.

[62]J. Gan, S. Li, Y. Zhai, and C. Liu, “3d convolutional neural network based on face anti-spoofing,” in *2017 2nd International Conference on Multimedia and Image Processing (ICMIP)*, March 2017, pp. 1–5.

[63]O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, “On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing,” in *2018 International Conference on Biometrics (ICB)*, Feb 2018, pp. 75–81.

[64]X. Zhang, X. Hu, M. Ma, C. Chen, and S. Peng, “Face spoofing detection based on 3d lighting environment analysis of image pair,” in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 2995–3000.

[65]A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, “The replay-mobile face presentation- attack database,” in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–7.

[66]J. Zhou, K. Shu, P. Liu, J. Xiang, and S. Xiong, “Face anti-spoofing based on dynamic color texture analysis using local directional number pattern,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan 2021, pp. 4221–4228.

[67]D. Wen, H. Han, and A. K. Jain, “Face spoof detection with image distortion analysis,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, April 2015.

[68]W. Sun, Y. Song, H. Zhao, and Z. Jin, “A face spoofing detection method based on domain adaptation and lossless size adaptation,” *IEEE Access*, vol. 8, pp. 66 553–66 563, 2020.

[69]E. M. Rudd, M. G ünther, and T. E. Boult, “Paraph: Presentation attack rejection by analyzing polarization hypotheses,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016, pp. 171–178.

[70]H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 578–593, 2020.
[71]W. Yin, Y. Ming, and L. Tian, "A face anti-spoofing method based on optical flow field," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, Nov 2016, pp. 1333–1337.
[72]P. P. P. Linn and E. C. Htoon, "Face anti-spoofing using eyes movement and cnn-based liveness detection," in
2019 International Conference on Advanced Information Technologies (ICAIT), Nov 2019, pp. 149–154. [73]S. Zhu, X. Lv, X. Feng, J. Lin, P. Jin, and L. Gao, "Plenoptic face presentation attack detection," *IEEE Access*,
vol. 8, pp. 59 007–59 014, 2020.
[74]L. Lv, Y. Xiang, X. Li, H. Huang, R. Ruan, X. Xu, and Y. Fu, "Combining dynamic image and prediction ensemble for cross-domain face anti-spoofing," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 2550–2554.
[75]S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and
Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 577–594.
[76]S. Jia, C. Hu, G. Guo, and Z. Xu, "A database for face presentation attack using wax figure faces," in *New Trends in Image Analysis and Processing – ICIAP 2019*, M. Cristani, A. Prati, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham: Springer International Publishing, 2019, pp. 39–47.
[77]A. K. Singh, P. Joshi, and G. C. Nandi, "Face recognition with liveness detection using eye and mouth move- ment," in *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014)*, July 2014, pp. 592–597.
[78]R. H. Vareto, A. Marcia Saldanha, and W. R. Schwartz, "The swax benchmark: Attacking biometric systems with wax figures," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 986–990.
[79]X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietik äinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 4244–4249.
[80]Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "Nas-fas: Static-dynamic central difference network search for face anti-spoofing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 9, pp. 3005–3023, Sep. 2021.
[81]S.-Q. Liu, X. Lan, and P. C. Yuen, "Temporal similarity analysis of remote photoplethysmography for fast 3d mask face presentation attack detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020, pp. 2597–2605 [82]Y. B. Reeba and R. Shanmugalakshmi, "Spoofing face recognition," in *2015 International Conference on Ad- vanced Computing and Communication Systems*, Jan 2015, pp. 1–5.
[83]G. Botelho de Souza, D. F. da Silva Santos, R. Gonçalves Pires, J. P. Papa, and A. N. Marana, "Efficient width- extended convolutional neural network for robust face spoofing detection," in *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, Oct 2018, pp. 230–235.
[84]D. P érez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, "Deep anomaly detection for gen- eralized face anti-spoofing," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Work- shops (CVPRW)*, June 2019, pp. 1591–1600.
[85]A. Ali, S. Hoque, and F. Deravi, "Biometric presentation attack detection using stimulated pupillary move- ments," in *9th International Conference on Imaging for Crime Detection and Prevention (ICDP-2019)*, Dec 2019, pp. 80–85.
[86]I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1713–1723, July 2017.
[87]K. Karthik and B. R. Katika, "Identity independent face anti-spoofing based on random scan patterns," in *Pattern Recognition and Machine Intelligence*, B. Deka, P. Maji, S. Mitra, D. K. Bhattacharyya, P. K. Bora, and S. K. Pal, Eds. Cham: Springer International Publishing, 2019, pp. 3–12.
[88]Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 557–575.
[89]Y. A. U. Rehman, L.-M. Po, M. Liu, Z. Zou, W. Ou, and Y. Zhao, "Face liveness detection using convolutional-features fusion of real and deep network generated face images," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 574–582, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320319300641

[90]J. Xia, Y. Tang, X. Jia, L. Shen, and Z. Lai, “Latent spatial features based on generative adversarial networks for face anti-spoofing,” in *Biometric Recognition*, Z. Sun, R. He, J. Feng, S. Shan, and Z. Guo, Eds. Cham: Springer International Publishing, 2019, pp. 240–249.

[91]H. Hao, M. Pei, and M. Zhao, “Face liveness detection based on client identity using siamese network,” in
Pattern Recognition and Computer Vision, Z. Lin, L. Wang, J. Yang, G. Shi, T. Tan, N. Zheng, X. Chen, and
Y. Zhang, Eds. Cham: Springer International Publishing, 2019, pp. 172–180.

[92]S. Komkov and A. Petiushko, “Advhat: Real-world adversarial attack on arcface face id system,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan 2021, pp. 819–826.

[93]R. B. Hadiprakoso, H. Setiawan, and Girinoto, “Face anti-spoofing using cnn classifier amp; face liveness detection,” in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, Nov 2020, pp. 143–147.

[94]M. Killio ğlu, M. Taşkiran, and N. Kahraman, “Anti-spoofing in face recognition with liveness detection using pupil tracking,” in *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Jan 2017, pp. 000 087–000 092.

[95]N. Alsufyani, A. Ali, S. Hoque, and F. Deravi, “Biometric presentation attack detection using gaze alignment,” in *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, Jan 2018, pp. 1–8.

[96]J. Dave, A. Khan, B. Gupta, A. Gangwar, and S. Suman, “Human-computer interaction methodology to attain face liveness detection,” in *2021 2nd International Conference for Emerging Technology (INCET)*, May 2021, pp. 1–4.

[97]L. Spinoulas, M. E. Hussein, D. Geissb ühler, J. Mathai, O. G. Almeida, G. Clivaz, S. Marcel, and W. Ab- dalmageed, “Multispectral biometrics system framework: Application to presentation attack detection,” *IEEE Sensors Journal*, vol. 21, no. 13, pp. 15 022–15 041, July 2021.

[98]L. Omar and I. Ivrissimtzis, “Designing a facial spoofing database for processed image attacks,” in *7th Interna- tional Conference on Imaging for Crime Detection and Prevention (ICDP 2016)*, Nov 2016, pp. 1–6.

[99]Z. Ji, H. Zhu, and Q. Wang, “Lfhog: A discriminative descriptor for live face detection from light field image,” in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 1474–1478.

[100]A. F. Ebihara, K. Sakurai, and H. Imaoka, “Efficient face spoofing detection with flash,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2021.

[101]R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch, “Face presentation attack detection by exploring spec- tral signatures,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 672–679.

[102]A. Jourabloo, Y. Liu, and X. Liu, “Face de-spoofing: Anti-spoofing via noise modeling,” in *Computer Vision*
– ECCV 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 297–315.

[103]K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, “Face anti-spoofing via disentangled representation learning,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 641–657.

[104]Y. Wang, X. Song, T. Xu, Z. Feng, and X.-J. Wu, “From rgb to depth: Domain transfer network for face anti- spoofing,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4280–4290, 2021.

[105]Q. Ji, S. Xu, X. Chen, S. Zhang, and S. Cao, “A cross domain multi-modal dataset for robust face anti-spoofing,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan 2021, pp. 4309–4316.

[106]H. Kaur and P. Khanna, “Random slope method for generation of cancelable biometric features,” *Pattern Recognition Letters*, vol. 126, pp. 31–40, 2019, robustness, Security and Regulation Aspects in Current Biometric Systems. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S016786551830059 X

[107]H. B. Alwan and K. R. Ku-Mahamud, “Cancellable face biometrics template using alexnet,” in *Applied Comput- ing to Support Industry: Innovation and Technology*, M. I. Khalaf, D. Al-Jumeily, and A. Lisitsa, Eds. Cham: Springer International Publishing, 2020, pp. 336–348.

[108]L. Ghammam, M. Barbier, and C. Rosenberger, “Enhancing the security of transformation based biometric template protection schemes,” in *2018 International Conference on Cyberworlds (CW)*, Oct 2018, pp. 316–323.

[109]L. C. O. Tiong, S. T. Kim, and Y. M. Ro, “Multimodal facial biometrics recognition: Dual-stream convolutional neural networks with multi-feature fusion layers,” *Image and Vision Computing*, vol. 102,

p. 103977, 2020. [Online].
Available:https://www.sciencedirect.com/science/article/pii/S0262885620301098
[110]F. CHEN, Y. SHANG, J. HU, and B. XU, "Few features attack to fool machine learning models through mask- based gan," in *2020 International Joint Conference on Neural Networks (IJCNN)*, July 2020, pp. 1–7.
[111]C. Park, D. Hong, and C. Seo, "An attack-based evaluation method for differentially private learning against model inversion attack," *IEEE Access*, vol. 7, pp. 124 988–124 999, 2019.
[112]Y. Alufaisan, M. Kantarcioglu, and Y. Zhou, "Robust transparency against model inversion attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2061–2073, Sep. 2021.
[113]S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, "Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes," in *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, Aug 2017, pp. 115–11 509.
[114]X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.
[115]M. Aladag, F. O. Catak, and E. Gul, "Preventing data poisoning attacks by using generative models," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Nov 2019, pp. 1–5.
[116]J. Chen, H. Zheng, M. Su, T. Du, C. Lin, and S. Ji, "Invisible poisoning: Highly stealthy targeted poisoning attack," in *Information Security and Cryptology*, Z. Liu and M. Yung, Eds. Cham: Springer International Publishing, 2020, pp. 173–198.
[117]A. S. Hashemi and S. Mozaffari, "Secure deep neural networks using adversarial image generation and training with noise-gan," *Computers & Security*, vol. 86, pp. 372–387, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740481930121X
[118]Y. Ma, K.-S. Jun, L. Li, and X. Zhu, "Data poisoning attacks in contextual bandits," in *Decision and Game Theory for Security*, L. Bushnell, R. Poovendran, and T. Bas¸ar, Eds. Cham: Springer International Publishing, 2018, pp. 186–204.
[119]A. Paudice, L. Mu ñoz-González, and E. C. Lupu, "Label sanitization against label flipping poisoning attacks," in
ECML PKDD 2018 Workshops, C. Alzate, A. Monreale, H. Assem, A. Bifet, T. S. Buda, B. Caglayan, B. Drury,
E. García-Martín, R. Gavaldà, I. Koprinska, S. Kramer, N. Lavesson, M. Madden, I. Molloy, M.-I. Nicolae, and
M. Sinn, Eds. Cham: Springer International Publishing, 2019, pp. 5–15.
[120]F. Tahmasebian, L. Xiong, M. Sotoodeh, and V. Sunderam, "Crowdsourcing under data poisoning attacks: A comparative study," in *Data and Applications Security and Privacy XXXIV*, A. Singhal and J. Vaidya, Eds. Cham: Springer International Publishing, 2020, pp. 310–332.
[121]E. E. B. Martinez, B. Oh, F. Li, and X. Luo, "Evading deep neural network and random forest classifiers by generating adversarial samples," in *Foundations and Practice of Security*, N. Zincir-Heywood, G. Bonfante,
M. Debbabi, and J. Garcia-Alfaro, Eds. Cham: Springer International Publishing, 2019, pp. 143–155.
[122]J. Wen, B. Z. H. Zhao, M. Xue, and H. Qian, "Palor: Poisoning attacks against logistic regression," in *Infor- mation Security and Privacy*, J. K. Liu and H. Cui, Eds. Cham: Springer International Publishing, 2020, pp. 447–460.
[123]J. Guo and C. Liu, "Practical poisoning attacks on neural networks," in *Computer Vision – ECCV 2020*,
A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 142–158.
[124]V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *Computer Security – ESORICS 2020*, L. Chen, N. Li, K. Liang, and S. Schneider, Eds.Cham: Springer International Publishing, 2020, pp. 480–501.
[125]R. Ning, J. Li, C. Xin, and H. Wu, "Invisible poison: A blackbox clean label backdoor attack to deep neural networks," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, May 2021, pp. 1–10.
[126]N. M üller, D. Kowatsch, and K. Böttinger, "Data poisoning attacks on regression learning and corresponding defenses," in *2020 IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC)*, Dec 2020, pp. 80–89.
[127]J. Wen, B. Z. H. Zhao, M. Xue, A. Oprea, and H. Qian, "With great dispersion comes greater resilience: Efficient poisoning attacks and defenses for linear regression models," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3709–3723, 2021.

[128]H. Chacon, S. Silva, and P. Rad, “Deep learning poison data attack detection,” in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov 2019, pp. 971–978.
[129]Y. Zhao, K. Xu, H. Wang, B. Li, and R. Jia, “Stability-based analysis and defense against backdoor attacks on edge computing services,” *IEEE Network*, vol. 35, no. 1, pp. 163–169, January 2021.
[130]B. Zhao and Y. Lao, “Resilience of pruned neural network against poisoning attack,” in *2018 13th International Conference on Malicious and Unwanted Software (MALWARE)*, Oct 2018, pp. 78–83.
[131]D. Cao, S. Chang, Z. Lin, G. Liu, and D. Sun, “Understanding distributed poisoning attack in federated learn- ing,” in *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, Dec 2019, pp. 233–239.
[132]N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.
[133]H. Rehman, A. Ekelhart, and R. Mayer, “Backdoor attacks in neural networks – a systematic evaluation on multiple traffic sign datasets,” in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg,
A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2019, pp. 285–300.
[134]Y. Xiong, F. Xu, S. Zhong, and Q. Li, “Escaping backdoor attack detection of deep learning,” in *ICT Systems Security and Privacy Protection*, M. Hölbl, K. Rannenberg, and T. Welzer, Eds. Cham: Springer International Publishing, 2020, pp. 431–445.
[135]R. Wang, G. Zhang, S. Liu, P.-Y. Chen, J. Xiong, and M. Wang, “Practical detection of trojan neural networks: Data-limited and data-free cases,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 222–238.
[136]Y. Liu, X. Ma, J. Bailey, and F. Lu, “Reflection backdoor: A natural backdoor attack on deep neural networks,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 182–199.
[137]J. Dumford and W. Scheirer, “Backdooring convolutional neural networks via targeted weight perturbations,” in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2020, pp. 1–9.
[138]Y. Li, J. Hua, H. Wang, C. Chen, and Y. Liu, “Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection,” in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, May 2021, pp. 263–274.
[139]X. Gong, Y. Chen, Q. Wang, H. Huang, L. Meng, C. Shen, and Q. Zhang, “Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment,” *IEEE Journal on Selected Areas in Communi- cations*, vol. 39, no. 8, pp. 2617–2631, Aug 2021.
[140]K. Alrawashdeh and S. Goldsmith, “Defending deep learning based anomaly detection systems against white- box adversarial examples and backdoor attacks,” in *2020 IEEE International Symposium on Technology and Society (ISTAS)*, Nov 2020, pp. 294–301.
[141]Y. Liu, Y. Xie, and A. Srivastava, “Neural trojans,” in *2017 IEEE International Conference on Computer Design (ICCD)*, Nov 2017, pp. 45–48.
[142]S. Kolouri, A. Saha, H. Pirsiavash, and H. Hoffmann, “Universal litmus patterns: Revealing backdoor attackin cnns,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 298–307.
[143]X. Zhang, R. Gupta, A. Mian, N. Rahnavard, and M. Shah, “Cassandra: Detecting trojaned networks from adversarial perturbations,” *IEEE Access*, pp. 1–1, 2021.
[144]Z. Xiang, D. J. Miller, and G. Kesidis, “Detection of backdoors in trained classifiers without access to the training set,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2020.
[145]W. Wang, J. Sun, and G. Wang, “Visualizing one pixel attack using adversarial maps,” in *2020 Chinese Automa- tion Congress (CAC)*, Nov 2020, pp. 924–929.
[146]D. Chen, R. Xu, and B. Han, “Patch selection denoiser: An effective approach defending against one-pixel attacks,” in *Neural Information Processing*, T. Gedeon, K. W. Wong, and M. Lee, Eds. Cham: Springer International Publishing, 2019, pp. 286–296.
[147]S. Huang, W. Peng, Z. Jia, and Z. Tu, “One-pixel signature: Characterizing cnn models for backdoor detection,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 326–341.
[148]H. Kwon, Y. Kim, H. Yoon, and D. Choi, “One-pixel adversarial example that is safe for friendly deep neural networks,” in *Information Security Applications*, B. B. Kang and J. Jang, Eds. Cham: Springer International Publishing, 2019, pp. 42–54.
[149]F. Croce and M. Hein, “A randomized gradient-free attack on relu networks,” in *Pattern Recognition*, T. Brox,
A. Bruhn, and M. Fritz, Eds. Cham: Springer International Publishing, 2019, pp. 215–227.

[150]X. Feng, H. Yao, W. Che, and S. Zhang, “An effective way to boost black-box adversarial attack,” in *MultiMedia Modeling*, Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, Eds. Cham: Springer International Publishing, 2020, pp. 393–404.
[151]M. Zhang, H. Li, X. Kuang, L. Pang, and Z. Wu, “Neuron selecting: Defending against adversarial examples in deep neural networks,” in *Information and Communications Security*, J. Zhou, X. Luo, Q. Shen, and Z. Xu, Eds. Cham: Springer International Publishing, 2020, pp. 613–629.
[152]W. Wan, J. Chen, and M.-H. Yang, “Adversarial training with bi-directional likelihood regularization for visual classification,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 785–800.
[153]J. Xu, Z. Cai, and W. Shen, “Using fgsm targeted attack to improve the transferability of adversarial example,” in *2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE)*, Dec 2019, pp. 20–25.
[154]N. Mani, M. Moh, and T.-S. Moh, “Towards robust ensemble defense against adversarial examples attack,” in
2019 IEEE Global Communications Conference (GLOBECOM), Dec 2019, pp. 1–6.
[155]A. P. Norton and Y. Qi, “Adversarial-playground: A visualization suite showing how adversarial examples fool deep learning,” in *2017 IEEE Symposium on Visualization for Cyber Security (VizSec)*, Oct 2017, pp. 1–4.
[156]S. Chhabra, A. Agarwal, R. Singh, and M. Vatsa, “Attack agnostic adversarial defense via visual imperceptible bound,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan 2021, pp. 5302–5309.
[157]J. Hayes and G. Danezis, “Learning universal adversarial perturbations with generative models,” in *2018 IEEE Security and Privacy Workshops (SPW)*, May 2018, pp. 43–49.
[158]S. Sharmin, P. Panda, S. S. Sarwar, C. Lee, W. Ponghiran, and K. Roy, “A comprehensive analysis on adversarial robustness of spiking neural networks,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, pp. 1–8.
[159]J. Guo, W. Ji, and Y. Li, “Generative networks for adversarial examples with weighted perturbations,” in *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Nov 2019, pp. 778–784.
[160]J. Lu, T. Issaranon, and D. Forsyth, “Safetynet: Detecting and rejecting adversarial examples robustly,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 446–454.
[161]Y. Zhang, W. Ruan, F. Wang, and X. Huang, “Generalizing universal adversarial attacks beyond additive per- turbations,” in *2020 IEEE International Conference on Data Mining (ICDM)*, Nov 2020, pp. 1412–1417.
[162]A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh, “Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition,” in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Oct 2018, pp. 1–7.
[163]Y. Deng and L. J. Karam, “Universal adversarial attack via enhanced projected gradient descent,” in *2020 IEEE International Conference on Image Processing (ICIP)*, Oct 2020, pp. 1241–1245.
[164]Y. Chu, X. Yue, Q. Wang, and Z. Wang, “Secureas: A vulnerability assessment system for deep neural network based on adversarial examples,” *IEEE Access*, vol. 8, pp. 109 156–109 167, 2020.
[165]H. Wang, G. Li, X. Liu, and L. Lin, “A hamiltonian monte carlo method for probabilistic adversarial attack and learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.
[166]T. Amada, S. P. Liew, K. Kakizaki, and T. Araki, “Universal adversarial spoofing attacks against face recogni- tion,” in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Aug 2021, pp. 1–7.
[167]A. Guesmi, I. Alouani, M. Baklouti, T. Frikha, and M. Abid, “Sit: Stochastic input transformation to defend against adversarial attacks on deep neural networks,” *IEEE Design Test*, pp. 1–1, 2021.
[168]L. Yu, T. Deng, W. Zhang, and Z. Zeng, “Stronger adversarial attack: Using mini-batch gradient,” in *2020 12th International Conference on Advanced Computational Intelligence (ICACI)*, Aug 2020, pp. 364–370.
[169]U. Sabeel, S. S. Heydari, H. Mohanka, Y. Bendhaou, K. Elgazzar, and K. El-Khatib, “Evaluation of deep learn- ing in detecting unknown network attacks,” in *2019 International Conference on Smart Applications, Commu- nications and Networking (SmartNets)*, Dec 2019, pp. 1–6.
[170]E. Lejeune, “Geometric stability classification: Datasets, metamodels, and adversarial attacks,” *Computer- Aided Design*, vol. 131, p. 102948, 2021. [Online].
Available:https://www.sciencedirect.com/science/article/pi i/S001044852030141X
[171]X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, “Face anti-spoofing: Model matters, so does data,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 3502–3511.

[172]X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton, “A methodology for formalizing model-inversion attacks,” in *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, June 2016, pp. 355–370.

[173]M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, “Manipulating machine learning: Poisoning attacks and countermeasures for regression learning,” in *2018 IEEE Symposium on Security and Privacy (SP)*, May 2018, pp. 19–35.

[174]R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 3–18.

[175]M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[176]M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, “Privacy in pharmacogenetics: An end-to- end case study of personal{ized war}farin dosing,” in *23rd {USENIX } Security Symposium (USENIX Security 14)*, 2014, pp. 17–32.

[177]T. Kubota, K. Yoshida, M. Shiozaki, and T. Fujino, “Deep learning side-channel attack against hardware implementations of aes,” *Microprocessors and Microsystems*, p. 103383, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141933120305408

[178]P. Wasnik, K. B. Raja, R. Raghavendra, and C. Busch, “Presentation attack detection in face biometric systems using raw sensor data from smartphones,” in *2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, Nov 2016, pp. 104–111.

[179]Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, D. Samai, S. E. Bekhouche, A. Ouafi, F. Dor- naika, A. Taleb-Ahmed, L. Qin, F. Peng, L. Zhang, M. Long, S. Bhilare, V. Kanhangad, A. Costa-Pazo,
E. Vazquez-Fernandez, D. Perez-Cabo, J. J. Moreira-Perez, D. Gonzalez-Jimenez, A. Mohammadi, S. Bhat- tacharjee, S. Marcel, S. Volkova, Y. Tang, N. Abe, L. Li, X. Feng, Z. Xia, X. Jiang, S. Liu, R. Shao, P. C. Yuen,
W. R. Almeida, F. Andalo, R. Padilha, G. Bertocco, W. Dias, J. Wainer, R. Torres, A. Rocha, M. A. Angeloni,
G. Folego, A. Godoy, and A. Hadid, “A competition on generalized software-based face presentation attack detection in mobile scenarios,” in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Oct 2017, pp. 688–696.

[180]C.-X. Wang, S.-Y. Zhao, X.-S. Wang, M. Luo, and M. Yang, “A neural network trojan detection method based on particle swarm optimization,” in *2018 14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, Oct 2018, pp. 1–3.

[181]M. Qin, W. Hu, X. Wang, D. Mu, and B. Mao, “Theorem proof based gate level information flow tracking for hardware security verification,” *Computers & Security*, vol. 85, pp. 225–239, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404819300975

[182]H. Kaur and P. Khanna, “Privacy preserving remote multi-server biometric authentication using cancelable biometrics and secret sharing,” *Future Generation Computer Systems*, vol. 102, pp. 30–41, 2020. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S0167739X18330553

[183]X. Zheng, L. Xie, H. Chen, and C. Song, “Performance analysis of consensus-based distributed system under fasle data injection attacks,” in *Communications and Networking*, H. Gao, Z. Feng, J. Yu, and J. Wu, Eds. Cham: Springer International Publishing, 2020, pp. 483–497.

[184]H. A. Al-Hamid and S. K. M. M. Rahman, “Securing photos in the cloud using decoy photo gallery,” in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, March 2017, pp. 816–822.

[185]D. F. Smith, A. Wiliem, and B. C. Lovell, “Binary watermarks: a practical method to address face recogni- tion replay attacks on consumer mobile devices,” in *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015)*, March 2015, pp. 1–6.

[186]M. A. Saleem, S. H. Islam, S. Ahmed, K. Mahmood, and M. Hussain, “Provably secure biometric- based client–server secure communication over unreliable networks,” *Journal of Information Security and Applications*, vol. 58, p. 102769, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S2214212621000181

[187]Y. Jie, K.-K. R. Choo, M. Li, L. Chen, and C. Guo, “Tradeoff gain and loss optimization against man-in-the- middle attacks based on game theoretic model,” *Future Generation Computer Systems*, vol. 101, pp. 169–179, 2019. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S0167739X18315541

[188]D. Cole, S. Newman, and D. Lin, “A new facial authentication pitfall and remedy in web services,” *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2021.

[189]Z. Ali, M. S. Hossain, G. Muhammad, I. Ullah, H. Abachi, and A. Alamri, “Edge-centric multimodal authentication system using encrypted biometric templates,” *Future Generation Computer Systems*, vol. 85, pp. 76–87, 2018. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S0167739X17328741

[190]Z. Zuo, X. Cao, and Y. Wang, “Security control of multi-agent systems under false data injection attacks,” *Neurocomputing*, vol. 404, pp. 240–246, 2020. [Online]. Available:https://www.sciencedirect.com/science/arti cle/pii/S0925231220307505

[191]H. Li, J. Zhang, and X. He, “Design of data-injection attacks for cyber-physical systems based on kullback–leibler divergence,” *Neurocomputing*, vol. 361, pp. 77–84, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219308173

[192]F. Younis and A. Miri, “Using honeypots in a decentralized framework to defend against adversarial machine- learning attacks,” in *Applied Cryptography and Network Security Workshops*, J. Zhou, R. Deng, Z. Li, S. Ma- jumdar, W. Meng, L. Wang, and K. Zhang, Eds. Cham: Springer International Publishing, 2019, pp. 24–48.

[193]K. Durkota, V. Lis ý, B. Bošanský, C. Kiekintveld, and M. Pěchouček, “Hardening networks against strategic attackers using attack graph games,” *Computers & Security*, vol. 87, p. 101578, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404819300689

[194]J. S. Abbasi, F. Bashir, K. N. Qureshi, M. Najam ul Islam, and G. Jeon, “Deep learning- based feature extraction and optimizing pattern matching for intrusion detection using finite state machine,” *Computers & Electrical Engineering*, vol. 92, p. 107094, 2021. [Online]. Available: https: //www.sciencedirect.com/science/article/pii/S0045790621001038

[195]S. Huang and K. Lei, “Igan-ids: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks,” *Ad Hoc Networks*, vol. 105, p. 102177, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570870519311035

[196]S.-J. Bu and S.-B. Cho, “Genetic algorithm-based deep learning ensemble for detecting database intrusion via insider attack,” in *Hybrid Artificial Intelligent Systems*, H. Pérez García, L. Sánchez González, M. Castejón Li- mas, H. Quintián Pardo, and E. Corchado Rodríguez, Eds. Cham: Springer International Publishing, 2019, pp. 145–156.

[197]A. Javed and M. Akhlaq, “On the approach of static feature extraction in trojans to combat against zero-day threats,” in *2014 International Conference on IT Convergence and Security (ICITCS)*, Oct 2014, pp. 1–5.

[198]M. A. Olivero, A. Bertolino, F. J. Dom ´ınguez-Mayo, M. J. Escalona, and I. Matteucci, “Digital persona portrayal: Identifying pluridentity vulnerabilities in digital life,” *Journal of Information Security and Applications*, vol. 52, p. 102492, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S2214212619308014

[199]A. Sepas-Moghaddam, F. Pereira, and P. L. Correia, “Light field-based face presentation attack detection: Reviewing, benchmarking and one step further,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1696–1709, July 2018.

[200]S. Purnapatra, N. Smalt, K. Bahmani, P. Das, D. Yambay, A. Mohammadi, A. George, T. Bourlai, S. Marcel,
S. Schuckers, M. Fang, N. Damer, F. Boutros, A. Kuijper, A. Kantarci, B. Demir, Z. Yildiz, Z. Ghafoory,
H. Dertli, H. K. Ekenel, S. Vu, V. Christophides, L. Dashuang, Z. Guanghao, H. Zhanlong, L. Junfu, J. Yufeng, S. Liu, S. Huang, S. Kuei, J. M. Singh, and R. Ramachandra, “Face liveness detection competition (livdet-face)
- 2021,” in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Aug 2021, pp. 1–10.

[201]U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, “Face recognition systems under morphing attacks: A survey,” *IEEE Access*, vol. 7, pp. 23 012–23 026, 2019.

[202]F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, “Adversarial attacks against face recognition: A comprehen- sive study,” 2021.

[203]Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, “A survey on security threats and defensive techniques of machine learning: A data driven view,” *IEEE Access*, vol. 6, pp. 12 103–12 117, 2018.

[204]Y. Liu, A. Mondal, A. Chakraborty, M. Zuzak, N. Jacobsen, D. Xing, and A. Srivastava, “A survey on neural trojans,” in *2020 21st International Symposium on Quality Electronic Design (ISQED)*, March 2020, pp. 33–39.

[205]F. V. Massoli, F. Carrara, G. Amato, and F. Falchi, “Detection of face recognition adversarial attacks,” *Computer Vision and Image Understanding*, vol. 202, p. 103103, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314220301296

[206]F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, “Adversarial attacks against face recognition: A comprehen- sive study,” *IEEE Access*, vol. 9, pp. 92 735–92 756, 2021.
[207]Y. He, G. Meng, K. Chen, X. Hu, and J. He, “Towards security threats of deep learning systems: A survey,”
IEEE Transactions on Software Engineering, pp. 1–1, 2020.
[208]X. Lin and C.-T. Li, “Enhancing sensor pattern noise via filtering distortion removal,” *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 381–385, March 2016.
[209]Y. Li, Y. Wang, and D. Li, “Privacy-preserving lightweight face recognition,” *Neurocomputing*, vol. 363, pp.
212–222, 2019. [Online].
Available:https://www.sciencedirect.com/science/article/pii/S0925231219310045 [210]H. Lee, J. Kim, S. Ahn, R. Hussain, S. Cho, and J. Son, “Digestive neural networks: A novel defense strategy against inference attacks in federated learning,” *Computers & Security*, vol. 109, p. 102378, 2021. [Online].
Available:https://www.sciencedirect.com/science/article/pii/S0167404821002029
[211]C. E. Lee, L. Zheng, Y. Zhang, V. L. L. Thing, and Y. Yu Chu, “Towards building a remote anti-spoofing face authentication system,” in *TENCON 2018 - 2018 IEEE Region 10 Conference*, Oct 2018, pp. 0321–0326.
[212]S. Patil, K. Tajane, and J. Sirdeshpande, “Enhancing security and privacy in biometrics based authentication system using multiple secret sharing,” in *2015 International Conference on Computing Communication Control and Automation*, Feb 2015, pp. 190–194.
[213]N. I. of Standards and Technology, Measuring strength of authentication. [Online]. Available: https:
//www.nist.gov/system/files/nstic-strength-authentication-discussion-draft.pdf.
[214]I. O. for Standardization, 2016. [Online]. Available: https://webstore.iec.ch/preview/info isoiec30107-3%7Be d1.0%7Den.pdf
[215]C. C. for Cyber Security, “Cyber threat and cyber threat actors,”: https://cyber.gc.ca/en/guidance/cyber-threat-a nd-cyber-threat-actors, 2021.
[216]E. Cybersecurity, “Enisa threat landscape,”https://www.enisa.europa.eu/topics/threat-risk-management/risk-m anagement/current-risk/risk-management-inventory/rm-isms/framework, 2021.
[217]D. Manky, “Cybercrime as a service: A very modern business,” *Computer Fraud & Security*, vol. 2013, p. 9–13, 06 2013.

10 List of Figures

11 List of Tables