The Alan Turing Institute

Technical Briefing

Threat Models for Face Recognition Systems: Taxonomy of Threats

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 Alan Turing Institute'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 Institute 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 Institute 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 and artificial intelligence, 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. Acknowledging a growing trend to adopt Facial Recognition Systems (FRS) in many user identification processes, this paper draws on a systematic review of attacks to articulate a systems architecture and taxonomy of threats and provide the context needed to identify and model risks to these systems.

It is part of a body of resources and guidance developed in consultation with governments, humanitarian organisations and the industry stakeholders that are advancing digital identity systems.

2 Executive Summary

Intuitively, weakly protected Facial Recognition Systems (FRS) would open the doors to various crimes and sabotage. These systems are increasingly deployed by major industries and governments, making their way into both foundational national identity systems and the functional systems that facilitate access to both private and public services, including border crossings, online commerce and more.

As governments and society develop reliance on digital identity technologies, such as face recognition, it is important that we recognise measures to be taken that can attest to their trustworthiness. Turing researchers are examining them 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.

This paper draws on a systematic review of attacks on facial recognition systems and their potential to compromise these pillars to articulate a new architecture and taxonomy of threats, advancing understanding of the vulnerabilities and threats that are evolving with the deployment of these systems. We elaborate a threat model identifying the most crucial elements in FRS to present a threat taxonomy categorised by whether they relate to the computing resources, the model employed, and/ or the system's input for authentication. We then present an Authentication Reference Architecture (ARA) and incorporate all threats present at each stage of the authentication process, describing and mapping 13 areas of attack in total. Here we elaborate more on the registration process compared to prior work. Reference data associated with this process is usually deemed to be clean and trusted, an assumption which could lead to the overlooking of severe flaws that compromise security, privacy, and the functional operations that underpin accountability for reliable and ethical access to services. Our analysis illustrates that this stage is effectively targeted in a number of adversarial attacks to compromise the

quality and resolution of the biometric samples collected or the models for creating reference templates from these samples. These include morphological attacks, which are growing in incidence and sophistication. Our ARA also links recently discovered attacks to the processing stages in user identification systems incorporating the registration stage.

These attacks and the vulnerabilities they exploit could severely hinder trust in FRS, leading to a reluctance in their adoption. In producing an attack taxonomy and a new reference architecture, we aim to increase confidence in the adoption of FRS and identify priority research areas for the prevention of attacks as part of future work.

3 Acronyms

FRS Facial Recognition Systems.

GAN Generative Adversarial Network.

IoC Indicator of Compromise.

NIST National Institute of Standards and Technology.

OSI Open Systems Interconnection.

POC Proof of Concept.

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses. ROC Receiver Operating Characteristics.

SPN Sensor Pattern Noise

PETS Privacy Enhancing Technologies

4 Existing Methods for Threat Modelling

In order to properly design FRS, we have to identify what can be targeted. Threat models concern analysing representations of a system to highlight concerns about security and privacy characteristics [1]. Threat modelling approaches have emerged in the early 2000s as a means to systematically identify and evaluate security requirements [2]. Following recent surveys [3– 5] we summarise some of the most prominent. The majority of the methods are mnemonic, i.e., the threat categories in question are coded in the method's name.

Attack Trees [6] to model threats are essentially diagrams that depict attacks on a system in tree form. The tree root is the goal for the attack, and the leaves are ways to achieve that goal. Each goal is represented as a separate tree. Thus, the system threat analysis produces a set of attack trees. Usually, it takes a few iterations of decomposing the goal to build the tree. Once all leaf nodes are identified, markers of possibility can be assigned. These values should be

assigned only after relevant research on the step is done. Goals can be accomplished in multiple ways.

Microsoft developed STRIDE that identifies five threat categories: spoofing, tampering, repudiation, disclosure, denial of service, and elevation. This method suffers from excessive time and resource consumption, as the number of threats can grow rapidly as a system increases in complexity [7, 8]. This tool was eventually replaced by Microsoft Threat Modelling Tool (TMT) [9], which is based on defining security requirements, creating an application diagram, identifying threats, mitigating threats, and validating that threats have been mitigated.

PASTA A life cycle model named **Process for Attack Simulation and Threat Analysis** [10] has as its main components: define objectives, define technical scope, application decomposition, threat analysis, vulnerability and weaknesses analysis, and attack modelling, risk and impact analysis. This method elevates the threat modelling process to a strategic level by involving key decision makers and requiring security input from operations, governance, architecture, and development. It is regarded as a risk-centric framework, facilitating analysis from the attacker perspective and revealing the potential for harm.

LINDDUN (Linkability, Identifiability, Non-Repudiation, Detectability, Disclosure of Information, Unawareness, and Non-Compliance) is a threat modelling method that focusses on privacy concerns and can be used for data security [11]. The main elements are: define a dataflow diagram, map privacy threats to diagram elements, identify threat scenarios, prioritise threats, elicit mitigation strategies, and select corresponding PETS.

The Open Web Application Security Project (OWASP) [12] is a software-focused threat model. OWASP follows the Threat Modelling Manifesto [1] which is based on answering four key questions: What are we working on? What can go wrong? What are we going to do about it? Did we do a good job? The model has the following components: threats agents, attacks vectors, security weaknesses, security controls, technical impacts, and business impacts.

The Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) [13] methodology was one of the first created specifically for information security threat modelling. OCTAVE threat modelling methodology is focused on assessing organisational risks that may result from breached data assets. Its main components are: build asset-based threat profiles, identify infrastructure vulnerability, and develop a security strategy and plans.

Trike [8, 14] was created as a security audit framework from risk management and defensive perspective. It is focused on satisfying the security auditing process from information risk management requirements. It provides a risk-based approach with a unique implementation, and risk modelling process. The requirements model ensures the assigned level of risk for each asset is acceptable by the stakeholders.

The Common Vulnerability Scoring System (CVSS) models the principal characteristics of a vulnerability and produces a numerical score reflecting its severity [15]. CVSS provides users of the method with a common and standardised scoring system within different cyber and cyber-physical platforms. The metrics are defined via some expert knowledge. The CVSS consists of

three metric groups: base (attack features assessment), temporal (response details), and environmental (confidentiality and integrity assessment).

The Visual, Agile, and Simple Threat VAST model is based on the generation of two sub-models: one application and the second for operations. The first sub-model uses process flow diagrams that represent the architectural point of view, while the latter focusses on the attackers' point of view based on the diagrams. One important feature of VAST is the scalability and usability that allows it to be adopted in large organisations throughout the entire infrastructure to produce actionable and reliable results for different stakeholders.

Persona non Grata (PnG) focuses on the motivations and skills of human attackers. It characterises users as archetypes that can misuse the system and forces analysts to view the system from an unintended use point of view [16]. It can help to visualise threats from the counterpart side, which can be helpful in the early stages of threat modelling.

Security Cards is a model that focuses on identifying unusual and complex attacks [17]. It means to answer basic questions about the potential threats: by whom? Why might the system be attacked? what assets are of interest? and how can these attacks be implemented? This method uses a deck of 42 cards to facilitate threat discovery activities: human impact (9 cards, e.g. financial well-being), adversary's motivations (13 cards, e.g. malice, money), adversary resources (11 cards, expertise, tools), and adversary's methods (9 cards, e.g. physical, cover-up).

Some threat models combine more than one structure to create a new one:

The Hybrid Threat Modelling Method (hTMM) [8, 18] consists of a combination of SQUARE (Security Quality Requirements Engineering Method), Security Cards**,** and Persona No Grata (PnG)**.** The targeted characteristics of the method include no false positives, no overlooked threats, a consistent result regardless of who is doing the threat modelling, and cost-effectiveness. The main steps are: identify the system to be modelled, apply security cards, remove unlikely PnG, summarise the results using a support tool, and continue with a formal risk assessment method.

Comprehensive, Lightweight Application Security Process (CLASP) [19] is an activity-driven, role-based set of process components guided by formalised best practices. CLASP is designed to help software development teams build security into the early stages of existing and new-start software development life cycles in a structured, repeatable, and measurable way. CLASP comprises 24 activities to ensure security via a simple three-tiered architecture. CLASP is conducive to iterative refinement, and it has metrics for evaluation. It comprises four main key tasks: identify system roles and resources, categorise resources into abstractions, identify resource interactions through the lifetime of the system, and for each category, specify mechanisms for addressing each core security service.

Quantitative Threat Modelling Method (QTMM) [20] is a hybrid method of Attack Trees, STRIDE, and CVSS. The first step of the QTMM is to build attack trees for the five threat categories of STRIDE. This activity shows the dependencies among attack categories and low-level component attributes. Next, the CVSS method is applied, and scores are calculated for the

components in the tree. An additional goal is to generate attack ports for individual components to illustrate activities that can pose a risk. The scoring assists with the process of performing a system risk assessment. If an attack port is dependent on a component root node with a high-risk score, that attack port also has a high-risk score and has a high probability of being executed.

Other recent methods adapt previous methods to create more flexible or robust models. Only recently, the European Union Agency for Cybersecurity (ENISA) proposed **Risk Management and Information Security Management Systems RM/ISMS** [21]. The framework is mainly based on OCTAVE nevertheless, when necessary, structural elements that emanate from other perceptions of risk management and risk assessment are incorporated as parts of wider operational processes. Risk Management is a recurrent activity that deals with the analysis, planning, implementation, control, and monitoring of implemented measurements and the enforced security policy. By contrast, Risk Assessment is executed at discrete time points (e.g. once a year, on-demand, etc.). The ISMS Framework is meant to implement the appropriate measurements to eliminate or minimise the impact that various security related threats and vulnerabilities might have on an organisation. It will enable implementing the desirable qualitative characteristics of the services offered by the organisation (i.e. availability of services, preservation of data confidentiality and integrity etc.).

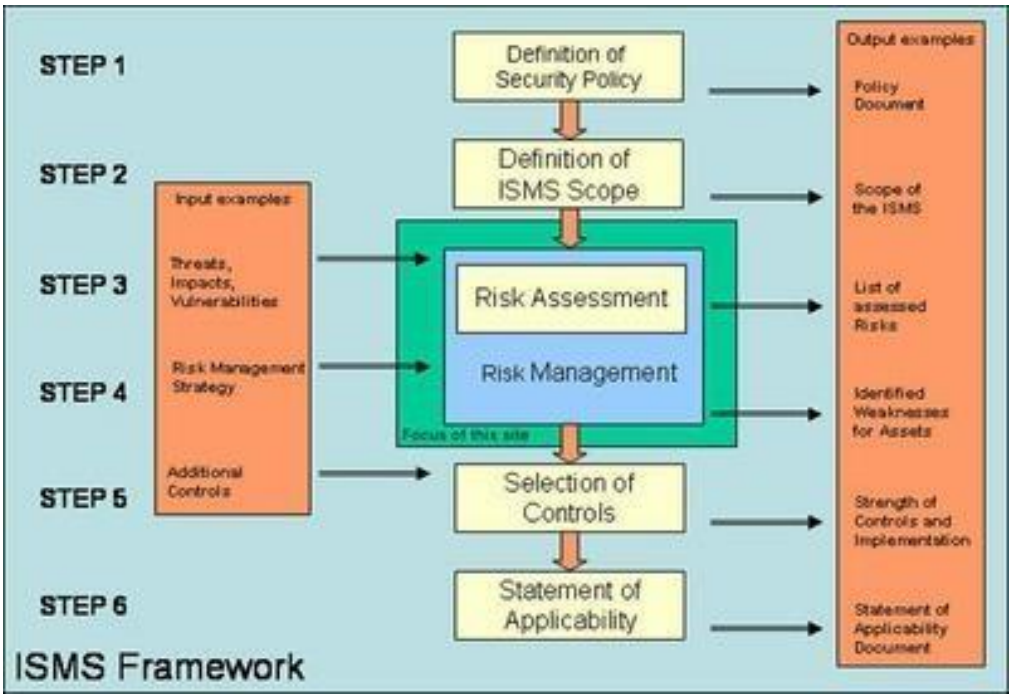

Fig 1: *ENISA ISMS Framework [21]*

The development of an ISMS framework requires the definition of the security policy, the definition of ISMS scope, risk assessment, risk management, selection of appropriate controls, and statement of applicability. Although the ISMS is a recurring process some steps are

lengthier than others. This is mainly because the establishment of a security policy and the definition of the ISMS scope are more often management and strategic issues while the Risk Management process is in the end an everyday operational concern.

5 Taxonomy of Threats Against Facial Recognition systems

Following the ENISA framework, we proceed to formulate the taxonomy of threats against FRS. The framework has at its core the requirement to identify threats, thus we proceed to elaborate on the subject based on the PRISMA literature review published in our companion Technical Briefing *Attacks Against Face Recognition Systems.* The following table summarises the threats towards FRS, grouping them into three main categories as described in the literature review. These are (1) computing resources, which are the technology needed to perform the task (2) models, which are the core of the signal processing to classify and authenticate users, and (3) input samples, which are the system's inputs. The table shows threats against FRS along with a brief description.

Table 1

Threats Towards FRS

Resource	High Level Threats	Threats	Threat details
	Physical	Fraud	Employees wrongdoings with personal or financial gains
		Sabotage	External actors' gains of the FRS failures, always intentional
		Vandalism	No real purpose but destroy assets
		Theft	Assets robbery, e.g., documents, hardware related to FRS.
		Information leaking	Sharing information with external actors
		Unauthorized physical access	Trespass facilities where the assets lie
	Unintentional	Information leaking	Information distribution unaware of the potential damage
		Device misuse	Information loss due to inappropriate device settings
		Non trusted source data	Taking decision based on non-reliable sources, e.g. datasets.
		Damage caused by third party	Result from penetration tests or white hacking activities

Computing/ Infrastructure		Information loss	Loss of integrity, in local or remote resources. Inadequate backups, moving files or losing records.
	Disasters	Natural	Floods, earthquakes, etc.
		Circumstantial	Fire, corrosion, dust.
	Failures	Systems	Caused by hardware, wrong design, scalability, implementation or complexity
		Communications	Infrastructure, communication protocols, compatibility.
		Power supply	Energy suppliers' failures.
	Outages	Resources loss	Unable to provide the system needs to operate.
		Personnel Absence	Strikes, or inadequate personnel planning.
		Internet	ISP unable to provide access
	Eavesdropping	Information Interception	Espionage and information stealing
		Source unveiling	Identify suppliers or operating sources
		Replay	Resending authentic messages to confuse the receiver
		Man in the Middle	Gain access to the data while transmitting
	Nefarious Activity	Identity theft	Have the credential to impersonate a victim

		Denial of Service	Make the service unavailable.
		Malware	Cause malfunctions or undesired behaviours at any stage of the system operation
		Social Engineering	Retrieve personal information from the system users
		Software, Hardware and Information manipulation	Manipulate the system components according to someone else's needs
		Data breach	Gain access to the system data needed to authenticate users
		Remote Access	Gain access the system to execute instructions arbitrarily
	Influence	Causative	Dataset manipulation
		Explorative	Misclassification

Models	Specificity	Targeted	Clearly identified sample to be accepted or rejected
		Indiscriminate	Cause rejection or acceptance for not specific sample(s)
	Violation	Availability	Increase false positives no matter the subject
		Integrity	Create false negatives of harmful samples
		Privacy	Retrieve information from the training set
Biometric Samples	Influence	Generative	Machine learning generated samples for authentication
		Spoofing	Authentic samples used from the attacker for authentication
		Intrusive	Tampered samples used for authentication
	Specificity	Targeted	Produce samples to authenticate a particular subject
		Indiscriminate	Authenticate no matter the subject

The table shows the most prominent threats against FRS. We use the ENISA threat taxonomy [21] as a baseline to identify threats in the context of FRS for authentication systems. To this end, we included the identified threats for FRS in the existing model. The rationale behind this is that the computing resource hosting the FRS faces the same threats as other authentication systems, e.g., man-in-the-middle, malware, frauds, etc. along with threats we identified for face biometrics, particularly. We present them as appended blocks in the table. These appended sections (summarised in the first column) correspond to threats towards the machine learning model that authenticate users, and the biometric samples the model receives as inputs. Following the ENISA threat landscape, the computing resource is deemed to be vulnerable to biometric sample threats as are other elements of the authentication system. biometric modalities [21]. The models and biometric samples' threats are specifically situated in the context of FRS. We follow our PRISMA approach to identify potential threats for those two categories. Threats towards machine learning models comprise the set in which the model in charge of authenticating users could be targeted. This set consists of three main branches, influence, specific, and violation as the table shows. The third group concerns the biometric samples which is the FRS input. This group comprises two main branches, influence and specificity as the table shows.

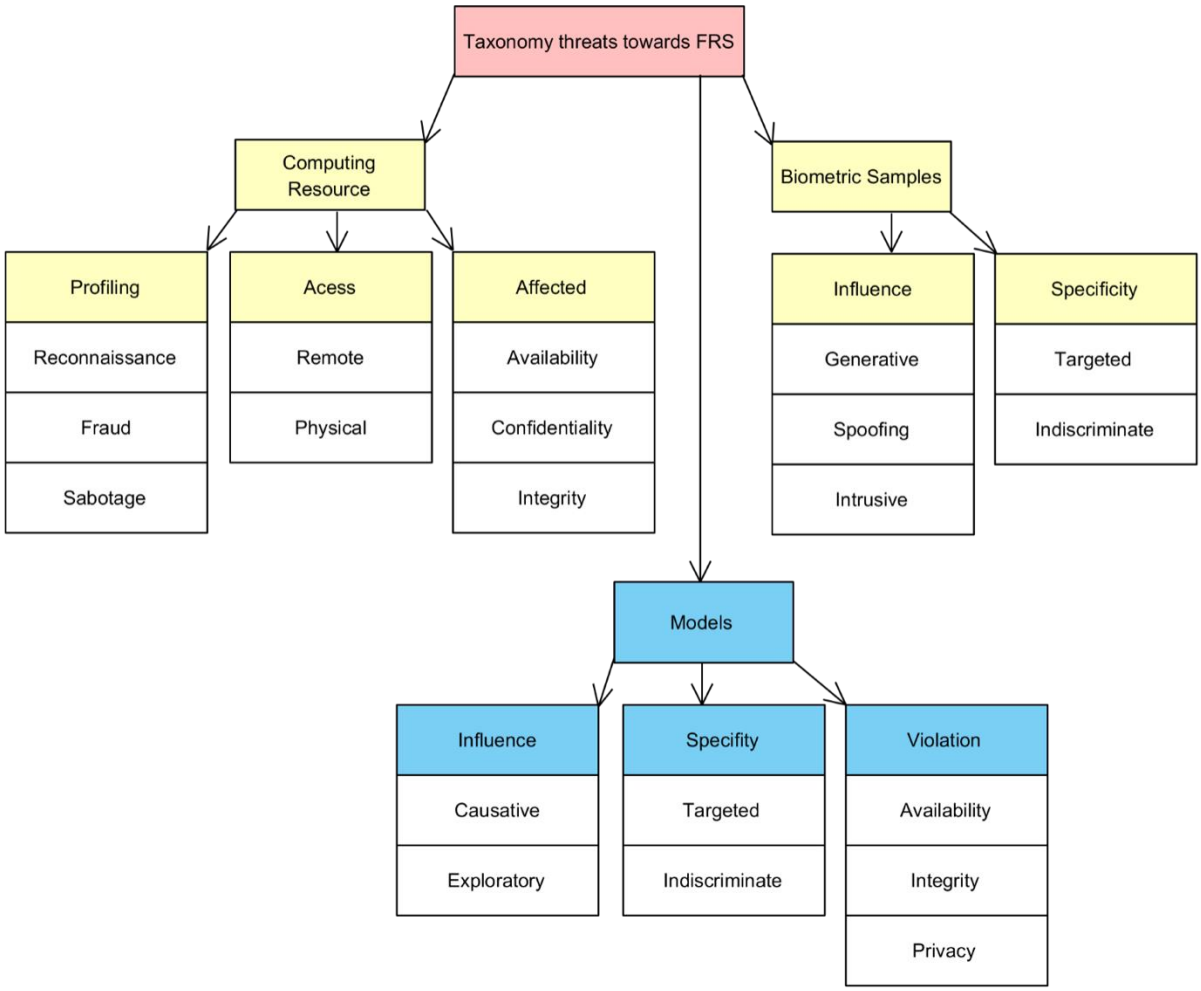

Fig. 2: *the taxonomy of security threats towards FRS from the high-level threats, see Table I.*

From the extended literature review, we identified the most prominent groups of the reported attacks in FRS and proceeded to identify the threats they represent. Ultimately these are linked to three main categories depending on their nature. The first group encapsulates the computing resource(s). We can deem this group as any other service for authenticating purposes; We associate this group to the existing recently reported subjects of the cyber threats. These represent by far the largest group. Due to the extensive work in the area, we did not elaborate extensively on this subject and employ the ENISA threat model as a baseline. For the remaining two groups that we elaborate more on, we can separate the system input (biometric sample) from the model used for the authentication. These are two independent units for which we can analyse threats separately. We created two groups at this point: the machine learning model and biometric sample input. We deem the machine learning model as the system's core. In this viewpoint, we analysed two aspects: the model itself and the data used to train the model. From the extended literature review, we identified the most prominent attacks and associate the corresponding threats to the models' group. Figure 2 shows this group as the Model Block that

comprises three sub-groups. The third group is the Biometric Samples. We follow a similar approach for the second group, from the extended literature review we identified the most persistent attacks and link them with the respective threats. During this process, we identified two main sub-groups we describe next. Figure 2 shows the last leaves from the Biometric Samples block. In summary, the taxonomy of security threats towards FRS comprises three different categories: the Computing Resources (yellow block) which is the infrastructure used to process the data and host all operations, Models (blue block) which define the core functions used to authenticate subjects using machine learning, and Biometric Samples (green block) which concerns the input data given to the system.

5.1 Computing Resources

We group the high-level threats to the computing resources into three groups: profiling, access and affected security:

Profiling: we can split this group into four main groups: (1) reconnaissance in which the victim is identified, (2) fraud where there is a personal or economic gain from the FRS user or service providers, (3) sabotage which implies destroying the FRS, and (4) eavesdropping where the attacker obtains important information from all the FRS actors.

Access: where threats concern gaining access to remote sources via the FRS authentication process, e.g., server logins and or physical access to equipment or facilities.

Affected security properties: we divide this group into three subgroups which concern (1) availability where FRS may not provide access to the desired resources, (2) confidentiality where the user data is exposed to the public and (3) integrity where the user data is subject to manipulation.

5.2 Biometric Samples

The second branch regards threats towards biometric samples, which we can divide into two main subgroups: influence and specificity.

Influence: we find three main subgroups (1) threats concerning generative attacks, where biometric samples are fabricated using machine learning models, (2) spoofing threats where authentic samples are used, and (3) intrusive where tampered samples are used somehow for impersonation.

Specificity: this group concerns two groups, (1) threats aimed at one subject and (2) indiscriminate, which is not aimed at one subject and manages the impersonation no matter the subject.

5.3 Models

The third branch regards threats towards the model, and it comprises three different groups: influence, specificity and violation.

Influence: This group comprises three main subgroups: (1) Causative threats. It means that adversaries can change the distribution of training data, which induces parameter changes of learning models when retraining, resulting in a significant decrease of the performance of classifiers in subsequent classification tasks. (2) Exploratory threats aim at causing misclassification with respect to adversarial samples or to uncover sensitive information from training data and learning models.

Violation: we can categorise these threats into three groups: (1) Integrity threats aim to increase the false negatives of existing classifiers when classifying harmful samples. (2) Availability threats will cause an increase of the false positives of classifiers concerning benign samples. (3) Privacy violation threats aim at obtaining sensitive and confidential information from training data and learning models.

Specificity: this group consists of two groups: (1) Targeted threats aiming to reduce the performance of classifiers on a particular sample or specific group of samples. (2) Indiscriminate threats cause the classifier to fail indiscriminately on a broad range of samples.

6 Authentication Reference Architecture

The authentication reference architecture details the FRS components. Prior work mainly focuses on the operation stage, placing particular importance on the presentation stage. Following the US National Institute of Standards and Technology (NIST) architecture, we elaborate an architecture incorporating the registration stage as an essential part of the FRS model. Figure 3 illustrates the proposed architecture. In this work, we address the existing vulnerabilities during the generation of the samples; as described earlier these can leave open doors to potential attacks [22, 23]. An extension of the NIST architecture comprises the upper left block of Figure 3. The architecture decomposes and facilitates analysis of the registration and authentication process separately to identify the potential threats. The Authentication Reference Architecture as proposed by NIST is represented in the green shaded block. The diagram depicts the system’s architecture in the presence of several attacks.

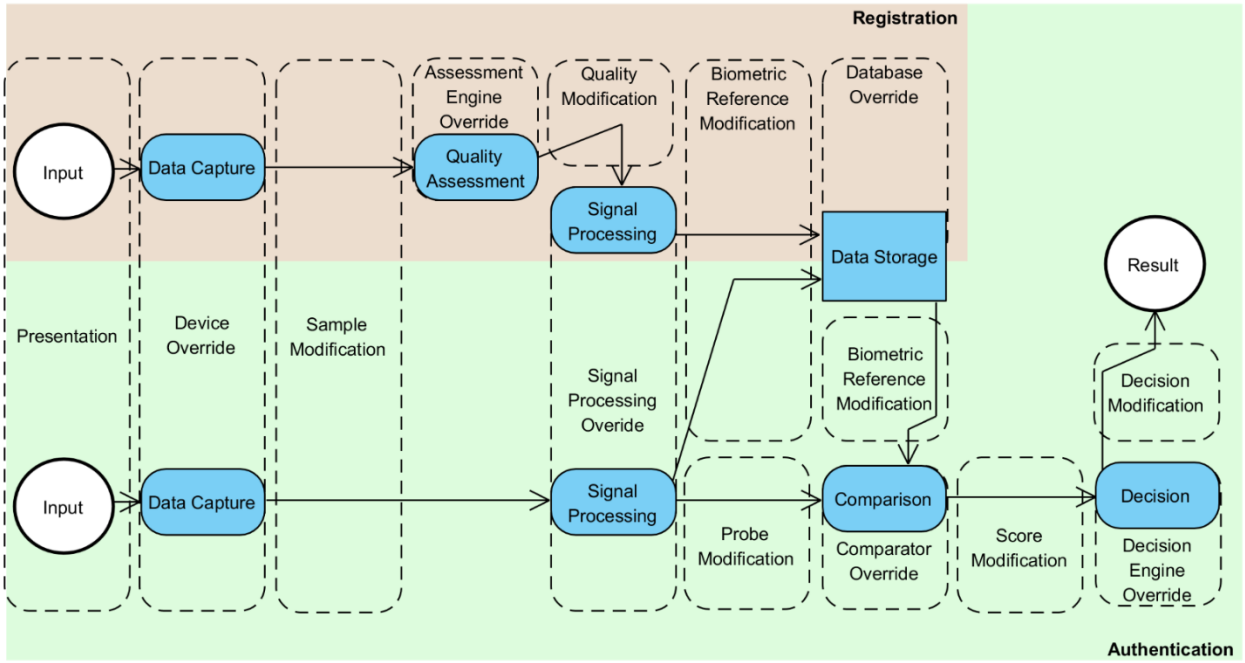

Fig. 3: *Proposed reference architecture.*

7 Thirteen Types of Attack

7.1 Presentation

Unfortunately, people's pictures are sometimes available on public websites without their knowledge [24, 25]. This fact is in part due to social media platforms' popularity. The attacker has plenty of resources to perform spoofing attacks, where they pretend to be the person in an image, which is particularly challenging when there is no controlled environment to perform the samples registration. Particularly for mobile devices and distributed authentication systems, no one inspects the biometric capture. Impersonation is easy to perform using available samples from public websites. With the attacker having access to user samples, the impersonation takes place using varied techniques, i.e., adversarial attacks [26–28], accessories [29,30], occlusions [31,32], video replay [33, 34], masks [35, 36], printouts [37, 38] and quality [39, 40] as detailed in our companion Technical Briefing *Attacks Against Face Recognition Systems*. Here the attacker's objective is generally to bypass the authentication no matter the subject and/or impersonation, however, creating distrust in FRS could also be an objective. A presentation attack could occur during the registration and authentication stages as Figure 3 depicts.

7.2 Device Override

The attacker may override the capture devices. In this case, FRS receive authentic or tampered samples never taken from the intended device [41] The attacker capabilities are backdoor access to the device used by the intended FRS, where the attackers' goals are impersonation and/or bypass. Device override threats are present both in the registration and authentication stages, as Figure 3 depicts.

7.3 Sample Modification

When performing the signal processing and quality assessment for the sample image, the transmission must be secure. This aspect is particularly essential when the device executes the sample capture and signal processing in place. Otherwise, mid-layer OSI attacks, e.g., man-in-the- middle, may take place and transmit modified samples [42, 43]. The attacker's capabilities are knowledge of the transmission channel and communication protocols. The goals are to impersonate and/or bypass authentication. Sample modification threats are present both in the registration and authentication stages, as Figure 3 depicts.

7.4 Signal Processing Override

It is necessary to process the samples by the signal processing method. This step involves extracting important information from the captured image which translates into a feature template. The model used in this stage may be corrupted in many forms, by the ways of poisoning [44–46], backdoors [47–49], and tested by adversarial samples [50–52], causing a complete manipulation of the processing. The attacker can perform an attack also by reversing the feature templates. These can be the result of elaborated tampering attacks such as inverse biometrics [53]. The attacker's capabilities are information about the feature extraction process or sample modelling used by the FRS. Where the goals are to impersonate and/or bypass the authentication signal processing threats are present both in the registration and authentication stages, as Figure 3 depicts.

7.5 Probe Modification

Using a similar strategy for insecure transmissions, the attacker may send unprocessed samples by the FRS during the first three system's stages. The attacker can send intercepted or tampered features templates and bypass the authentication [54, 55]. This, in practice, can be considered as a complete override of the signal processing stage. The attacker's capabilities

require knowledge of the transmission channel and communication protocols. Here the goals are to impersonate and/or bypass authentication.

7.6 Comparator Override

The algorithm in charge of deciding whether the processed sample matches an existing user may be targeted and modified via backdoor access, e.g. trojans. The attack may be reflected in extreme poor ROC's performance of the algorithm if the modification lets many non-authorised users bypass [56]. However, it can also be the case that the attacker has knowledge of the matching process and mounts attacks via the sample processing foundation. This can be achieved by finding very similar samples in feature spaces where the matching takes place [57]. The attacker's capabilities for both cases require knowledge of the matching process. The goals are to impersonate and/or bypass authentication.

7.7 Database Override

Properly storing the data is crucial to maintain integrity. If implemented successfully, the attacker may modify the data without compromising the entire FRS' security. In adding the additional database data, the signal processing model interprets it as authentic [58–60]. It could also be the case that already stored data is being modified [61–63]. However, more serious threats to the data being manipulated are represented by backdoors [64, 65]. The attacker's capabilities are knowledge of the transmission channel and communication protocols. The goals are to impersonate, inflect user rejection, and/or bypass the authentication.

7.8 Biometric Reference Modification

Similarly, the attacker may modify samples sent to the database during the transmission. This opens potential vulnerabilities to data injection [62] and data override [42]. The attacker's capabilities comprise knowledge of the transmission channel and communication protocols. The goals are impersonation, user rejection, and/or authentication bypass. Biometric reference modification threats are present both in the registration and authentication stages, as Figure 3 depicts.

7.9 Score Modification

Poorly protected systems could allow the attacker to interfere with the system's scoring. In this case, the comparison criteria are modified according to the attacker's needs to score fabricated samples differently. The attacker's capabilities are to access the scoring system, e.g., via a

backdoor [65]. Where the goals are impersonation, user rejection, authentication bypass, and/or distrust.

7.10 Decision Engine Override

The attacker could target the scores and/or thresholds to accept samples, manipulating the FRS to accept samples based on very weak decisions. This hinders the system's capabilities to discard fabricated samples after a respective comparison. The attacker's capabilities are to access the decision engine, e.g., via a backdoor [65]. The goals are impersonation, user rejection, authentication bypass, and/or distrust.

7.11 Decision Modification

The attacker can bypass the entire FRS directly by modifying the final decision as if the system did not exist. In this case, the input sample is irrelevant as the attacker can access the decision engine, e.g., via backdoors [65, 66]. The goals are impersonation, user rejection, authentication bypass, and/or distrust.

7.12 Assessment Engine Override

Low-quality samples pose a serious threat to the authentication stage [39, 40]. This is a particularly fertile ground for morph attacks [67–69] which recently have attracted attention by producing high-quality datasets for their mitigation [70]. This situation worsens in uncontrolled environments that lack robust quality assessment. Further, like any other stage of FRS, quality assessment can be avoided. Similarly, the assessment is under threat for weakly protected transmissions. The attacker's capabilities are knowledge of the transmission channel and communication protocols. The goals are to impersonate and/or bypass authentication. Assessment engine threats are present in the registration stage only, as Figure 3 depicts.

7.13 Quality Modification

Having similar foundations of previously described attacks, the attacker may compromise the image quality assessment whenever insecure transmission protocols are in place. In such situations, the assessment may not even take place. The attacker's capabilities are knowledge of the transmission channel and communication protocols. The goals are to impersonate and/or bypass authentication. Quality modification threats are present in the registration stage only, as Figure 3 depicts.

8 Summary

We presented a taxonomy of threats for FRS from an extensive and recent PRISMA literature review. Our model identifies three main targets of the FRS system. This comprises the common cyber threats faced by other authentication systems, and the model and sample threats exclusive to FRS. This taxonomy could inform the design of trustworthy systems with measures to mitigate the security, privacy and other threats elevated in this work. We group the most persistent attacks and dissect them based on their nature to present an architecture reference that advances the context needed for modelling current and developing threats to these systems. To this end, we elaborate on the registration stage as a cornerstone to mitigate the most persistent threat, the presentation attacks, countering common perception that reference datasets are deemed to be clean. Such an assumption opens doors to corrupted samples and low-quality images that may eventually compromise the whole system's security. Our extensive review helps to identify potential attacks at every system's component, including those could originate in the registration process or compromise its datasets. As part of the future work, we will tackle the second most persistent threat regarding models, which in combination with the presentation group could address the most prominent threats reported in the literature.

9 References

[1] Z. Braiterman, A. Shostack, J. Marcil, S. de Vries, I. Michlin, K. Wuyts, R. Hurlbut, B. S. Schoenfield, F. Scott, M. Coles, C. Romeo, A. Miller, I. Tarandach, A. Douglen, , and M. French, "Threat modeling manifesto," 2021. [online]. Available: http://www.threatmodelingmanifesto.org/

[2] M. Howard and S. Lipner, "The security development lifecycle," 2006.

[3] N. Shevchenko, T. A. Chick, P. O'Riordan, T. P. Scanlon, and C. Woody, "Threat modeling: a summary of available methods," Carnegie Mellon University Software Engineering Institute Pittsburgh United . . . , Tech. Rep., 2018.

[4] W. Xiong and R. Lagerström, "Threat modeling - a systematic literature review," Computers & Security, vol, 84, pp. 53–69, 2019. [online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404818307478

[5] D. Van Landuyt and W. Joosen, "A descriptive study of assumptions made in linddun privacy threat elicitation," in Proceedings of the 35th Annual ACM Symposium on Applied Computing, 2020, pp. 1280–1287.

[6] B. Schneier, "Attack trees," Dr. Dobb's journal, vol. 24, no. 12, pp. 21–29, 1999.

[7] Shostack, Threat modeling: Designing for security. John Wiley & Sons, 2014.

[8] N. R. Mead, F. Shull, K. Vemuru, and O. Villadsen, "A hybrid threat modeling method," Carnegie MellonUniversity-Software Engineering Institute-Technical Report-CMU/SEI-2018-TN-002, 2018. 4arXiv Template A P REPRINT

[9] M. Corporation, "Sdl threat modeling tool. security development lifecycle," 2021. [online]. Available: //www.microsoft.com/en-us/sdl/adopt/threatmodeling.aspx https:

[10] T. UcedaVelez, "Real world threat modeling using the pasta methodology," OWASP App Sec EU, 2012.

[11] K. Wuyts, D. Van Landuyt, A. Hovsepyan, and W. Joosen, "Effective and efficient privacy threat modeling through domain refinements," in Proceedings of the 33rd Annual ACM Symposium on Applied Computing, 2018, pp. 1175–1178.

[12] "Owasp threat modeling," 2021. [online]. Available: https://owasp.org/www- community/Threat_Modeling

[13] C. Alberts, A. Dorofee, J. Stevens, and C. Woody, "Introduction to the octave approach, software engineering institute," 2003.

[14] P. Saitta, B. Larcom, and M. Eddington, "Trike v. 1 methodology document [draft]," URL: http://dymaxion. org/trike/Trike v1 Methodology Documentdraft.

pdf, 2005.

[15] P. Mell, K. Scarfone, and S. Romanosky, “Common vulnerability scoring system,” IEEE Security & Privacy, vol. 4, no. 6, pp. 85–89, 2006.

[16] J. Cleland-Huang, “How well do you know your personae non gratae?” IEEE software, vol. 31, no. 4, pp. 28–31, 2014.

[17] T. Denning, B. Friedman, and T. Kohno, “The security cards: A security threat brainstorming toolkit,” Univ. of Washington, http://securitycards. cs. washington. edu, 2013.

[18] M. N., E. Hough, and S. T., “Security quality requirements engineering technical report,” 2005. [online]. Available: https://resources.sei.cmu.edu/library/asset-view.cfm?assetID=7657

[19] D. of Homeland Security, “Attack trees.” [online]. Available: http://web.cs.du.edu/~ramki/papers/attackGraphs/SchneierAttac kTrees.pdf

[20] B. Potteiger, G. Martins, and X. Koutsoukos, “Software and attack centric integrated threat modeling for quantitative risk assessment,” in Proceedings of the Symposium and Bootcamp on the Science of Security, 2016, pp. 99–108.

[21] E. U. A. F. Cybersecurity, “Enisa threat landscape,” 2021. [online]. Available: https://www.enisa.europa.eu/topics/threat-risk -management/risk-management/current-risk/risk- management-inventory/rm-isms/framework

[22] 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/pii/S001044852030141X

[23] 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.

[24] M. A. Olivero, A. Bertolino, F. J. Domínguez-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

[25] A. R. Shahid, N. Pissinou, S. Iyengar, and K. Makki, “Check-ins and photos: Spatiotemporal correlation-based location inference attack and defense in location-based social networks,” in 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), Aug 2018, pp. 1852–1857.

[26] 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.

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

[28] 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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, “Generalized face anti-spoofing by detecting pulse from face videos,” in 2016 23rd International Conference on Pattern Recognition (ICPR), Dec 2016, pp. 4244–4249. 5arXiv Template A P REPRINT

[36] 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.

[37] 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.

[38] X. Zhao, Y. Lin, and J. Heikkilä, “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.

[39] 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.

[40] 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 20, pp. 4309–4316.

[41] 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

[42] 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

[43] X. Zheng, L. Xie, H. Chen, and C. Song, “Performance analysis of consensus-based distributed system under false 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.

[44] 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]. https: //www.sciencedirect.com/science/article/pii/S016740481930121X

[45] 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. Başar, Eds. Cham: Springer International Publishing, 2018, pp. 186–204.

[46] A. Paudice, L. Muñoz-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.

[47] 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.

[48] 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 Communications, vol. 39, no. 8, pp. 2617–2631, Aug 2021.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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

[56] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic coun- termeasures," in Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 2015, pp. 1322–1333. 6arXiv Template A P REPRINT

[57] 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/S016786551830059X

[58] 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/article/pii/S0925231220307505

[59] 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/S0925231 219308173

[60] 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. Majumdar, W. Meng, L. Wang, and K. Zhang, Eds. Cham: Springer International Publishing, 2019, pp. 24–48.

[61] 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

[62] 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, pp169–179,2019. [onlinewww.sciencedirect.com/science/article/pii/S0167739X18315541

[63] 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.

[64] K. Durkota, V. Lisý, 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

[65] 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

[66] 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 Limas, H. Quintián

Pardo, and E. Corchado Rodríguez, Eds. Cham: Springer International Publishing, 2019, pp. 145–156.

[67] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. J. Veldhuis, L. Spreeuwers, M. Schils, D. Maltoni, 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.

[68] 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.

[69] 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.

[70] 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