IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 June 2024, accepted 28 July 2024, date of publication 7 August 2024, date of current version 30 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3439741

RESEARCH ARTICLE

A Comprehensive Risk Analysis Method for Adversarial Attacks on Biometric Authentication Systems

**SEONG HEE PARK, SOO-HYUN LEE, MIN YOUNG LIM,
PYO MIN HONG, AND YOUN KYU LEE**
Department of Computer Engineering, Hongik University, Seoul 04066, Republic of Korea

Corresponding author: Youn Kyu Lee (younkyul@hongik.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by Korean Government (MSIT) under Grant RS-2022-00165648, and in part by the 2024 Hongik University Research Fund.

ABSTRACT Recent threats to deep learning-based biometric authentication systems stem from adversarial attacks exploiting vulnerabilities in deep learning models. While existing studies extensively analyze the risk of such attacks, they primarily focus on isolated modules (e.g., liveness detectors or identity matchers) or specific adversarial attack types (e.g., evasion and poisoning attacks). In this paper, we introduce a novel approach that comprehensively assesses the risk of adversarial attacks by simulating multiple scenarios within biometric authentication systems. We identify the surfaces susceptible to adversarial attacks within these systems and devise scenarios that reflect the dependencies between modules. Moreover, we establish evaluation metrics to comprehensively assess the risk involved. Through a case study conducted on a real-world face recognition system, we successfully demonstrate the effectiveness of our approach. Our approach facilitates the systematic evaluation of the security of target biometric authentication systems against adversarial attacks. Ultimately, it enables the establishment of robust and proactive defense mechanisms.

INDEX TERMS Adversarial attack, adversarial attack scenarios, biometric authentication system, comprehensive risk analysis, deep learning.

I. INTRODUCTION

Recent biometric authentication systems employ deep learning-based authentication mechanisms to achieve high authentication accuracy [1], [2]. However, the security of these systems is threatened by adversarial attacks that exploit vulnerabilities in deep learning models [3]. Adversarial attacks lead to misclassification by subtly modifying data and deep learning algorithms or extract critical information by injecting malicious data into deep learning models. These attacks pose security threats to deep learning-based biometric authentication systems, including false authentication, access denial, and personal information theft.

To address these threats, a number of studies have analyzed the risks posed by adversarial attacks on biometric authentication systems [4]. Existing studies primarily focus on adversarial attacks targeting specific modules of biometric authentication systems [3], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. However, biometric authentication systems are composed of multiple modules (e.g., liveness detector and identity matcher) that interact with each other [18], [19]. The liveness detector determines the authenticity of the input biometric trait, while the identity matcher verifies its correspondence with registered users. For example, in a typical biometric authentication system, input data traverses through the liveness detector before reaching the identity matcher. In such cases, adversarial attacks targeting the identity matcher face constraints, as their manipulated input data must successfully pass through the liveness detector without being detected. Although the constraints of attack scenarios can vary depending on the dependencies between modules, existing studies often focus solely on

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Jin.

specific modules, neglecting the dependencies among them. This oversight may result in invalid vulnerability analyses of biometric authentication systems. Furthermore, existing studies primarily concentrate on particular types of adversarial attacks (e.g., evasion attacks, poisoning attacks, and exploratory attacks) [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [20]. However, this narrow focus may lead to biased analyses, overlooking other types of risks and making it challenging to quantitatively evaluate the risks associated with different attacks. Therefore, a comprehensive method is required, which systematically evaluates the risks of multiple types of adversarial attacks while considering the dependencies between modules in biometric authentication systems.

In this paper, we define adversarial attack scenarios in biometric authentication systems and introduce a systematic method for assessing the risk of adversarial attacks based on these scenarios. Our approach defines adversarial attack scenarios by considering the dependencies between modules in biometric authentication systems and enables a comprehensive evaluation of the risks from multiple perspectives through scenario-based adversarial attack simulations. Specifically, we identified seven adversarial attack surfaces within biometric authentication systems through a comprehensive analysis of existing systems and adversarial attacks. Furthermore, we defined a total of 12 adversarial attack scenarios for each attack surface, considering the dependencies between the system modules. Finally, to evaluate the risk of target systems based on those scenarios from multiple perspectives, we defined three evaluation metrics: attack vulnerability, attack execution difficulty, and attack defense availability. To validate the applicability of our approach, we conducted a case study using a face recognition system implemented with FaceNet [21] and CASIA-Webface [22].

Our approach facilitates the systematic assessment of biometric authentication system security against adversarial attacks, considering multiple potential attack scenarios and evaluating risks from various perspectives. This ultimately enables a comprehensive risk analysis of target biometric authentication systems and the establishment of proactive defense mechanisms against adversarial attacks. The contributions of this study are as follows: (1) Introducing a novel systematic method for assessing the risk of adversarial attacks on biometric authentication systems; (2) Defining various adversarial attack scenarios considering the dependencies between modules of biometric authentication systems; (3) Defining evaluation metrics to assess the risk of adversarial attacks from multiple perspectives; (4) Conducting a case study using real-world datasets on a face recognition system.

This paper is organized as follows: Section II presents related work, Section III describes the main approach, Section IV presents a case study, followed by a discussion in Section V, and finally, Section VI presents the conclusion.

II. RELATED WORK

A. CLASSIFICATION OF ADVERSARIAL ATTACKS

Adversarial attacks pose security threats by exploiting vulnerabilities in deep learning models, resulting in misclassification of input data or the leakage of critical information [23]. In biometric authentication systems, these attacks can be categorized based on criteria such as goal, type, and capability [4], [5], [24].

The goals of adversarial attacks on biometric authentication systems can be classified into three categories: integrity violation, availability violation, and privacy violation. Integrity violation aims to mimic a specific registered user within the target system. Availability violation disrupts users from accessing the target system. Privacy violation entails maliciously accessing the target system to extract data.

The types of adversarial attacks on biometric authentication systems can be categorized into evasion attacks, poisoning attacks, and exploratory attacks. Evasion attacks result in misclassification by injecting adversarial examples with subtle perturbations into deep learning models. Poisoning attacks disrupt the logic of deep learning models by manipulating the training process. Exploratory attacks extract critical information from deep learning models by analyzing the model's output for the attacker's input.

The capabilities of adversarial attacks on biometric authentication systems encompass both the training and testing phases. During the training phase, attacks manipulate the training dataset or algorithm of the deep learning model. These attack vectors are categorized into data modification and injection, which targets the data, and logic corruption, which targets the algorithm. Data modification and injection involve altering specific parts of the training dataset or injecting adversarial data into it. Logic corruption maliciously manipulates the training or inference algorithms of deep learning models. Conversely, during the testing phase, the attack vectors can be classified as either white-box or black-box. White-box attacks involve inputting adversarial data into the target model with knowledge of the model's information, such as algorithm, structure, and training parameters. Black-box attacks, on the other hand, input adversarial data into the target model without prior knowledge of the model's information. Generally, black-box attacks have lower success rates compared to white-box attacks.

To ensure robust protection for a target deep learning-based biometric authentication system against adversarial attacks, it's essential to implement a systematic risk assessment method that comprehensively addresses the diverse range of these attacks.

B. ADVERSARIAL ATTACK METHODS ON BIOMETRIC AUTHENTICATION SYSTEMS

In deep learning-based biometric authentication systems, vulnerabilities to adversarial attacks have been extensively reported [3], [25]. Fei et al. [3] introduced an adversarial attack method that enhances robustness to image

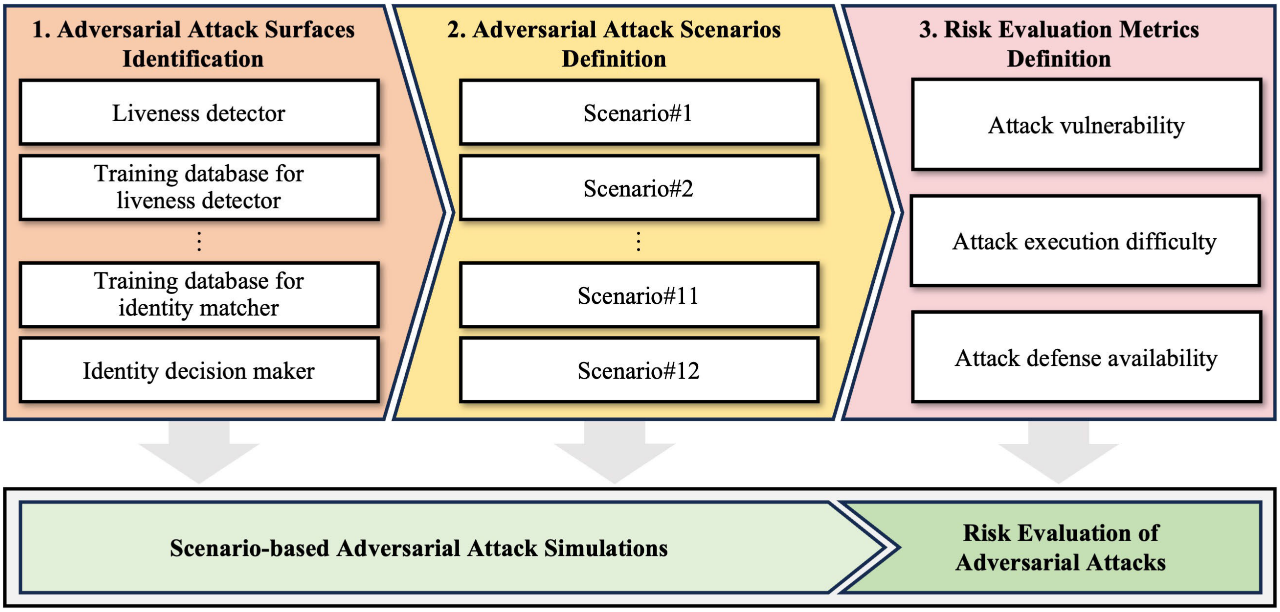
1. Adversarial Attack Surfaces Identification
Liveness detector
Training database for liveness detector
⋮
Training database for identity matcher
Identity decision maker
2. Adversarial Attack Scenarios Definition
Scenario#1
Scenario#2
⋮
Scenario#11
Scenario#12
3. Risk Evaluation Metrics Definition
Attack vulnerability
Attack execution difficulty
Attack defense availability
Scenario-based Adversarial Attack Simulations
Risk Evaluation of Adversarial Attacks

FIGURE 1. **Overview of the proposed approach for comprehensive risk evaluation of adversarial attacks in deep learning-based biometric authentication systems.**

transformations, such as flips and rotations, in fingerprint liveness detection. Their approach adds slight Gaussian noise and applies small-angle random rotations during each iteration. Yin et al. [26] proposed a makeup-based adversarial attack method aimed at the identity matcher of face recognition systems, incorporating eye shadow synthesis with perturbations onto facial images. Xue et al. [27] developed the Linear Offset based Poisoning Attack method (LOPA), exploiting the data update process of adaptive fingerprint authentication systems lacking a liveness detector. LOPA subtly infects the enrolled user's fingerprints in the system by introducing poisoning fingerprints resembling those of a specific user. Zhang et al. [28] proposed a generative model inversion attack, reconstructing the training dataset of deep learning models using a Generative Adversarial Network (GAN). They successfully exploited the identity matcher of face recognition systems. Given the variety of adversarial attack methods proposed for biometric authentication systems, an effective defense against these attacks requires a comprehensive risk assessment method that considers both the vulnerabilities of individual modules and their dependencies.

III. MAIN APPROACH

In this paper, we introduce a novel approach to assessing the risk of adversarial attacks on biometric authentication systems, based on potential attack scenarios, as shown in Fig. 1. We systematically identify various attack surfaces and define specific adversarial attack scenarios. Furthermore, we introduce three distinct evaluation metrics designed to comprehensively assess the risk of adversarial attacks inherent in biometric authentication systems from multiple perspectives. Based on these, our approach facilitates a comprehensive evaluation of the risks to the target system through scenario-based adversarial attack simulations and subsequent risk assessments.

A. BACKGROUND

Our approach is based on the typical structure of deep learning-based biometric authentication systems, with careful consideration of its practical applicability. Fig. 2 provides a depiction of a typical deep learning-based biometric authentication system [5], [18], [29], with the details of each module shown in Table 1. In this system, users' biometric traits are captured via sensors such as cameras and fingerprint scanners. Subsequently, the captured traits undergo sequential verification via two key modules: the liveness detector and the identity matcher. These modules commonly rely on deep learning models, which autonomously extract features from the input data and undergo training to generate predictive outputs.

This system primarily operates through two key processes: enrollment and authentication. Enrollment involves users registering their biometric traits in the system, following this sequence: (E1) Users input their biometric traits into the system via the sensor for a predefined number of times; (E2) The system digitizes the biometric traits z captured by the sensor and adds them to the training database for the identity matcher; (E3) The liveness detector undergoes training using a dataset comprising both real and fake biometric traits from the training database; (E4) The identity matcher undergoes training using a dataset containing biometric traits from registered users in the training database for the identity matcher, including the biometric traits z added in (E2).

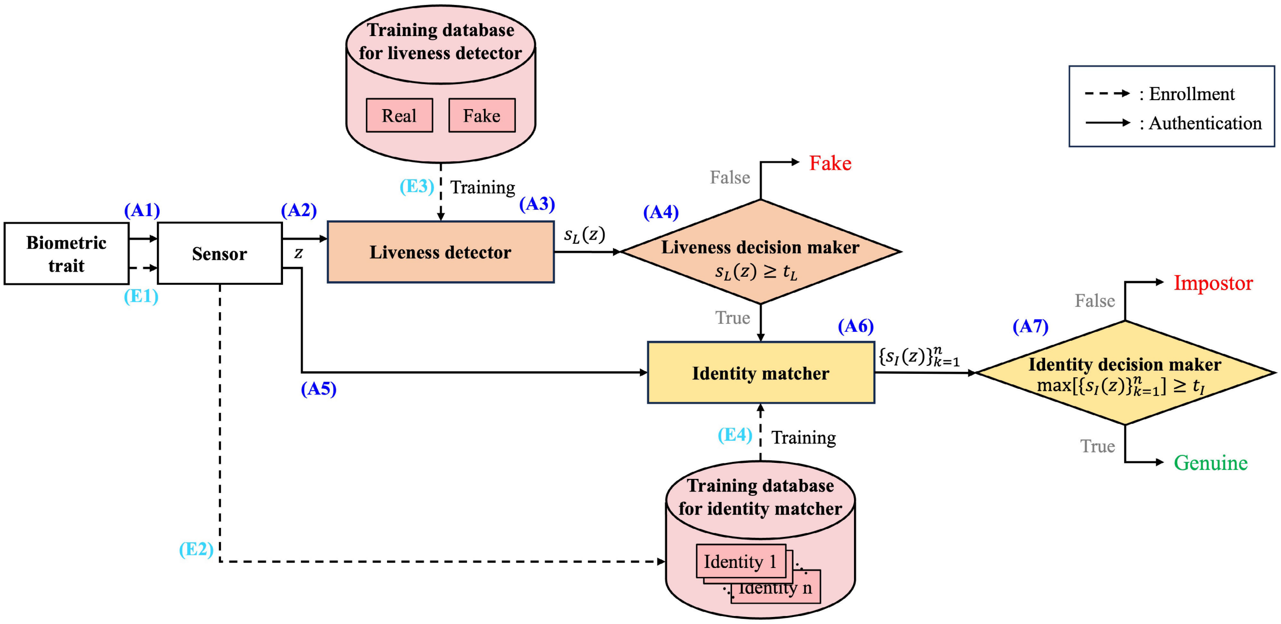

FIGURE 2. Diagram of a typical deep learning-based biometric authentication system, showing the enrollment sequence (E#) and authentication sequence (A#).

TABLE 1. Modules in typical deep learning-based biometric authentication system.

No.	Module	Description
1	Sensor	The module for acquiring the user's biometric trait into the biometric authentication system.
2	Liveness detector	The module for detecting whether the biometric trait input into the biometric authentication system has been tampered with. It is typically implemented based on a deep learning model and calculates the liveness score of the biometric trait digitized through a sensor.
3	Training database for liveness detector	The database for training the liveness detector in the biometric authentication system. It consists of both real and fake biometric traits.
4	Liveness decision maker	The module for determining whether the biometric trait input into the biometric authentication system is real or fake. It compares a pre-defined liveness threshold with the liveness score calculated from the liveness detector.
5	Identity matcher	The module for matching whether the biometric trait input into the biometric authentication system is a registered trait. It is typically implemented based on a deep learning model and calculates the identity score of the biometric trait digitized through a sensor.
6	Training database for identity matcher	The database for training the identity matcher in the biometric authentication system. It consists of the biometric traits of users registered in the biometric authentication system.
7	Identity decision maker	The module for determining whether the biometric trait input into the biometric authentication system is genuine or impostor. It compares a pre-defined identity threshold with the identity score calculated from the identity matcher.

On the other hand, authentication involves verifying a user's biometric input traits, following this sequence: (A1) Users input their biometric traits into the system via the sensor; (A2) The system digitizes the biometric trait z captured by the sensor and feeds it into the liveness detector to determine its authenticity; (A3) The liveness detector extracts the liveness feature of the input z and computes the liveness score $s_L(z)$; (A4) If $s_L(z)$ falls below the liveness threshold t_L set by the liveness decision maker, z is flagged as fake, resulting in access denial. Conversely, if $s_L(z)$ exceeds t_L, indicating authenticity, the process continues; (A5) The biometric trait z, having passed the liveness detector, is then forwarded to the identity matcher to verify its matching identity; (A6) The identity matcher extracts the identity feature of the input z and computes the identity score $[s_I(z)]_{k=1}^{n}$ for n identities trained by the identity matcher; (A7) If the highest score among the identity scores, denoted as $max([s_I(z)]_{k=1}^{n})$, falls below the identity threshold t_I set by the identity decision maker, z is identified as an impostor, leading to access denial. Conversely, if $max([s_I(z)]_{k=1}^{n})$ surpasses t_I, indicating authenticity, access is granted with the corresponding identity registered in the system.

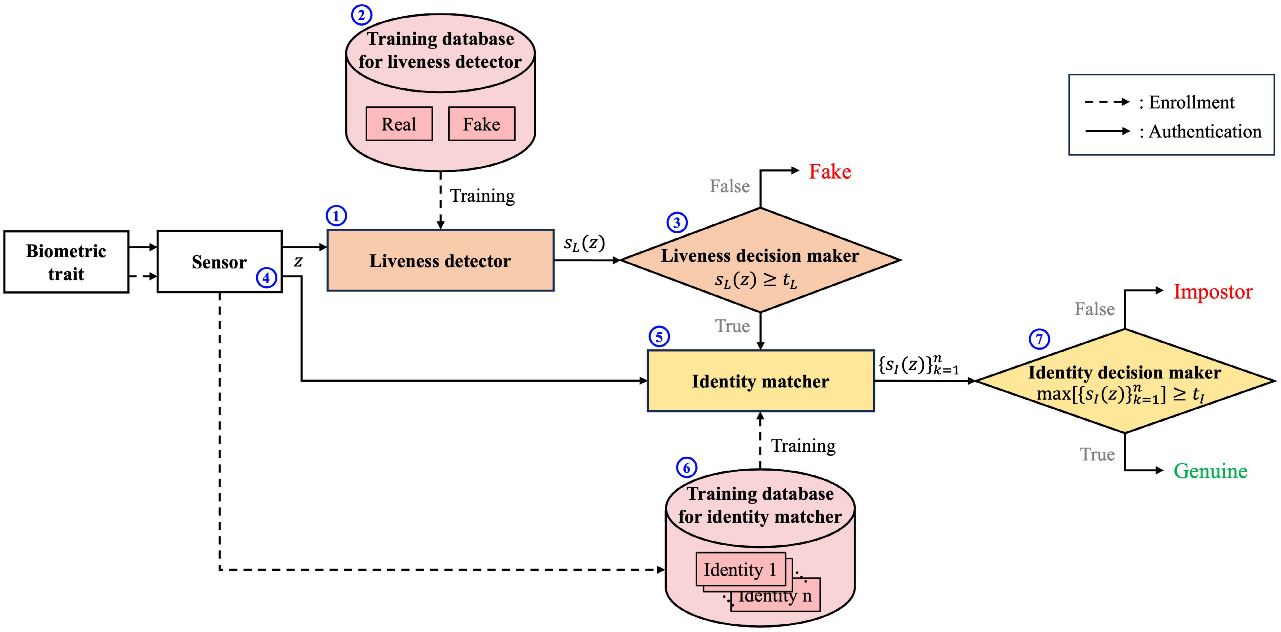

FIGURE 3. Adversarial attack surfaces where adversarial attacks can occur within typical deep learning-based biometric authentication systems.

For example, we assume a situation where attacker A attempts to impersonate user B by inputting a deepfake video—synthesizing B's face onto A's facial area—into the face recognition system. If the liveness detector detects the input video as manipulation, attacker A is denied access to the system. However, if the liveness detector does not identify the input video as manipulation, the deepfake video passes through the liveness detector and is forwarded to the identity matcher. Subsequently, the identity matcher checks whether the identity of the input video matches any identity registered in the system. If the identity matcher determines that none of the registered identities match the identity of the input video, attacker A is denied access to the system. Nevertheless, if the identity matcher matches the identity of the input video to user B, attacker A successfully impersonates user B and gains access to the system.

B. ADVERSARIAL ATTACK SURFACE

As depicted in Fig. 3, we identified seven adversarial attack surfaces within typical deep learning-based biometric authentication systems. Adversarial attack surfaces represent the points where adversarial attacks can occur within the target system, with each point potentially involving various attack scenarios. For instance, if an attacker gains access to the training database for the liveness detector, they could inject malicious data into the database (=data injection) or maliciously modify the existing data within the database (=data modification).

To identify the adversarial attack surfaces, we systematically analyzed existing research related to attack surfaces and adversarial attacks on biometric authentication systems. For this analysis, we conducted a keyword-based literature search for attack surfaces and adversarial attacks, focusing on papers published after 2015. We utilized search engines such as IEEE Explore, ACM Digital Library, Springer Link, and Google Scholar. Specifically, to analyze the attack surfaces in biometric authentication systems, we employed eight search queries by combining keywords related to the attack surface (e.g., attack surface and attack point) with those related to biometric authentication systems (e.g., biometric system, biometric authentication system, biometric process, and biometric authentication process). Similarly, to analyze adversarial attacks on biometric authentication systems, we utilized four search queries by combining 'adversarial attack' with keywords related to biometrics (e.g., biometric authentication, liveness, face, and fingerprint). Our search yielded a total of 253 papers related to attack surfaces in biometric authentication systems and 115 papers related to adversarial attacks within the same systems. The authors meticulously reviewed these papers, excluding those deemed less relevant to our study or the biometric authentication system. Ultimately, we curated a selection of 39 papers [5], [7], [9], [10], [11], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63] on attack surfaces and 59 papers [3], [6], [7], [8], [20], [30], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116] on adversarial attacks on biometric authentication systems.

TABLE 2. Adversarial attack scenarios within biometric authentication systems, based on the adversarial attack surfaces and the three criteria for categorizing adversarial attacks (i.e., goal, type, and capability).

No.	Adversarial Attack Surface	Goal	Type	Capability	
				Phase	Attack Vector
1	Liveness detector	Availability violation	Poisoning attack	Training phase	Logic corruption
2	Training database for liveness detector	Availability violation	Poisoning attack	Training phase	Data modification and injection
3	Liveness decision maker	Availability violation	Poisoning attack	Training phase	Logic corruption
4	Sensor-identity matcher	Integrity violation	Evasion attack	Testing phase	White-box
5	Sensor-identity matcher	Integrity violation	Evasion attack	Testing phase	Black-box
6	Sensor-identity matcher	Privacy violation	Exploratory attack	Testing phase	White-box
7	Sensor-identity matcher	Privacy violation	Exploratory attack	Testing phase	Black-box
8	Identity matcher	Integrity violation	Poisoning attack	Training phase	Logic corruption
9	Identity matcher	Availability violation	Poisoning attack	Training phase	Logic corruption
10	Training database for identity matcher	Availability violation	Poisoning attack	Training phase	Data modification and injection
11	Training database for identity matcher	Integrity violation	Poisoning attack	Training phase	Data modification and injection
12	Identity decision maker	Availability violation	Poisoning attack	Training phase	Logic corruption

We selected the adversarial attack surfaces by analyzing these papers and matched each with the modules within typical deep learning-based biometric authentication systems. Specifically, first, we composed an attack surface set by collecting attack surfaces from the selected papers. Then, we analyzed research related to adversarial attacks on biometric authentication systems and identified the attack surfaces involved in each attack. Through this process, we identified the following attack surfaces: deep learning model, decision maker, biometric database, and sensor-recognition model. Each attack surface was matched with the corresponding module within a typical deep learning-based biometric authentication system, as described in Section III-A. Since the deep learning model, decision maker, and database each consists of two elements (i.e., liveness detection and identity matcher), we matched them as follows: (1) deep learning model ↔ liveness detector and identity matcher, (2) decision maker ↔ liveness decision maker and identity decision maker, and (3) biometric database ↔ training database for liveness detector and training database for identity matcher. Additionally, as the recognition model corresponds to the identity matcher in this system, we matched sensor-recognition model ↔ sensor-identity matcher. As a result, we identified a total of seven adversarial attack surfaces: liveness detector, training database for liveness detector, liveness decision maker, sensor-identity matcher, identity matcher, training database for identity matcher, and identity decision maker.

C. ADVERSARIAL ATTACK SCENARIO

To evaluate the risk of adversarial attacks that may occur during the system's operation, we delineated 12 adversarial attack scenarios, considering the dependencies between modules within the biometric authentication system, as shown in iTable 2. These scenarios are based on the adversarial attack surfaces outlined in Section III-B, along with the three criteria for categorizing adversarial attacks (i.e., goal, type, and capability). Each attack scenario was formulated under the assumption of a successful attack without any intervention from the system. The details of each attack scenario are elaborated as follows:

(1) Scenario#1: In this scenario, the attacker's goal is an *availability violation*. Since the attacker can intervene in the *training phase* of the liveness detector, the adversarial attack surface is the *liveness detector*, and the attack type is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the liveness detector and performs *logic corruption* that modifies part of the training algorithm to consistently detect the user's input as fake. Consequently, the liveness detector consistently identifies all user-input biometric traits as fake, resulting in continuous denial of access to the system for the user.

(2) Scenario#2: In this scenario, the attacker's goal is an *availability violation*. Since the attacker can intervene in the *training phase* of the liveness detector, the adversarial attack surface is the *training database for the liveness detector*, and the attack type is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the liveness detector and performs *data modification and injection*, injecting malicious data into the training database for the liveness detector or directly modifying its data to consistently detect the user's input as fake. Consequently, the liveness detector consistently identifies all user-input biometric traits as fake, leading to continuous denial of access to the system for the user.

(3) Scenario#3: In this scenario, the attacker's goal is an *availability violation*. Since the attacker can intervene in the *training phase* of the liveness detector, the adversarial attack surface is the *liveness decision maker*, and the attack type

is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the liveness detector and performs *logic corruption* to modify the pre-defined liveness threshold of the liveness decision maker. The attacker adjusts the threshold to consistently detect the user's input as fake. Consequently, the liveness decision maker consistently identifies all user-input biometric traits as fake, leading to continuous denial of access to the system for the user.

(4) Scenario#4: In this scenario, the attacker's goal is an *integrity violation*. Since the attacker cannot intervene in the training phase of the biometric authentication system and can only input data through the sensor during the *testing phase*, the adversarial attack surface is the *sensor-identity matcher*, and the attack type is an *evasion attack*. Specifically, the attacker executes a *white-box* attack, inputting an adversarial biometric trait designed to consistently be recognized as a specific identity, utilizing information about the identity matcher of the biometric authentication system. Consequently, the adversarial biometric trait input by the attacker is detected as real by the liveness detector and then forwarded to the identity matcher. Subsequently, the identity matcher matches the adversarial biometric trait to a specific user, enabling the attacker to gain access to the system by impersonating that user.

(5) Scenario#5: In this scenario, the attacker's goal is an *integrity violation*. Since the attacker cannot intervene in the training phase of the biometric authentication system and can only input data through the sensor during the *testing phase*, the adversarial attack surface is the *sensor-identity matcher*, and the attack type is an *evasion attack*. Specifically, the attacker executes a *black-box attack*, inputting an adversarial biometric trait designed to consistently be recognized as a specific identity, without possessing information about the identity matcher of the biometric authentication system. Consequently, the adversarial biometric trait input by the attacker is detected as real by the liveness detector and then forwarded to the identity matcher. Subsequently, the identity matcher matches the adversarial biometric trait to a specific user, enabling the attacker to gain access to the system by impersonating that user.

(6) Scenario#6: In this scenario, the attacker's goal is a *privacy violation*. Since the attacker cannot intervene in the training phase of the biometric authentication system and can only input data through the sensor during the *testing phase*, the adversarial attack surface is the *sensor-identity matcher*, and the attack type is an *exploratory attack*. Specifically, the attacker executes a *white-box* attack, repeatedly inputting arbitrary biometric traits to analyze the output results, with information about the identity matcher of the biometric authentication system. Consequently, the biometric trait input by the attacker is detected as real by the liveness detector and then forwarded to the identity matcher. By analyzing the output of the identity matcher for the biometric trait, the attacker extracts the data trained by the identity matcher, resulting in a privacy violation in the system.

(7) Scenario#7: In this scenario, the attacker's goal is a *privacy violation*. Since the attacker cannot intervene in the training phase of the biometric authentication system and can only input data through the sensor during the *testing phase*, the adversarial attack surface is the *sensor-identity matcher*, and the attack type is an *exploratory attack*. Specifically, the attacker executes a *black-box* attack, repeatedly inputting arbitrary biometric traits to analyze the output results, without possessing information about the identity matcher of the biometric authentication system. Consequently, the biometric trait input by the attacker is detected as real by the liveness detector and then forwarded to the identity matcher. By analyzing the output of the identity matcher for the biometric trait, the attacker extracts the data trained by the identity matcher, resulting in a privacy violation in the system.

(8) Scenario#8: In this scenario, the attacker's goal is an *integrity violation*. Since the attacker can intervene in the *training phase* of the identity matcher, the adversarial attack surface is the *identity matcher*, and the attack type is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the identity matcher and performs *logic corruption* that modifies part of the training algorithm to detect the attacker's input as a specific identity consistently. Consequently, the attacker-input biometric trait is detected as real by the liveness detector and forwarded to the identity matcher. Subsequently, the identity matcher consistently matches the adversarial biometric trait to a specific user, enabling the attacker to gain access to the system by impersonating that user.

(9) Scenario#9: In this scenario, the attacker's goal is an *availability violation*. Since the attacker can intervene in the *training phase* of the identity matcher, the adversarial attack surface is the *identity matcher*, and the attack type is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the identity matcher and performs *logic corruption* that modifies part of the training algorithm to misclassify the user's input consistently. Consequently, the user-input biometric trait is detected as real by the liveness detector and forwarded to the identity matcher. The identity matcher fails to match the user-input biometric trait with the authenticated user, leading to continuous denial of access to the system for the user.

(10) Scenario#10: In this scenario, the attacker's goal is an *availability violation*. Since the attacker can intervene in the *training phase* of the identity matcher, the adversarial attack surface is the *training database for the identity matcher*, and the attack type is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the identity matcher and performs *data modification and injection*, injecting malicious data into the training database for the identity matcher or directly modifying its data to consistently misclassify the user's input. Consequently, the user-input biometric trait is detected as real by the liveness detector and forwarded to the identity matcher. The identity matcher fails to match the

TABLE 3. Types of information required for executing the adversarial attack.

No.	Type	Description
1	Model architecture [24]	The architecture of the deep learning model in the target system.
2	Parameters [4]	The parameters updated during the training process of the deep learning model in the target system.
3	Training data [4]	The data used to train the deep learning model in the target system.
4	Output result [117]	The output resulting from input to the deep learning model in the target system.
5	Target label [24]	The target label indicating the misclassification results from the deep learning model in the target system desired by the attacker.
6	Training algorithm [24]	The training algorithm used to train the deep learning model in the target system.

user-input biometric trait with the authenticated user, leading to continuous denial of access to the system for the user.

(11) Scenario#11: In this scenario, the attacker's goal is an *integrity violation*. Since the attacker can intervene in the *training phase* of the identity matcher, the adversarial attack surface is the *training database for the identity matcher*, and the attack type is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the identity matcher and performs *data modification and injection*, injecting malicious data into the training database for the identity matcher or directly modifying its data to consistently detect the attacker's input as a specific identity. Consequently, the attacker-input biometric trait is detected as real by the liveness detector and forwarded to the identity matcher. Subsequently, the identity matcher matches the biometric trait to a specific user, enabling the attacker to gain access to the system by impersonating that user.

(12) Scenario#12: In this scenario, the attacker's goal is an *availability violation*. Since the attacker can intervene in the *training phase* of the identity matcher, the adversarial attack surface is the *identity decision maker*, and the attack type is a *poisoning attack*. Specifically, the attacker intervenes in the training phase of the identity matcher and performs *logic corruption* that modifies the pre-defined identity threshold of the identity decision maker. The attacker adjusts the threshold to consistently detect the user's input as an imposter. Consequently, the user-input biometric trait is detected as real by the liveness detector and forwarded to the identity matcher. Subsequently, the identity decision maker identifies the user-input biometric trait as an impostor, leading to continuous denial of access to the system for the user.

D. RISK EVALUATION METRIC

We defined evaluation metrics to assess the risk of adversarial attacks on biometric authentication systems from various perspectives. Adversarial attacks on these systems differ in the required information and defense strategies based on the goal, type, and capability of each attack. To comprehensively evaluate the risk of adversarial attacks, we established three metrics: (1) attack vulnerability, (2) attack execution difficulty, and (3) attack defense availability. The detailed descriptions are as follows.

1) ATTACK VULNERABILITY

Attack vulnerability is a metric used to quantitatively assess the susceptibility of the biometric authentication system to each adversarial attack scenario. This metric relies on the attack success rate, considered one of the representative metrics for evaluating the risk of adversarial attacks [3], [8], [10], [11], [16]. It is determined by the number of successful attacks on the testing dataset used for attack simulation in each scenario. The formula is as follows:

$$(Attack\ vulnerability) = \frac{(The\ number\ of\ data\ with\ successful\ attack)}{(The\ total\ number\ of\ testing\ data)} \times 100 \quad (1)$$

High attack vulnerability suggests that the attack scenario used in the test has a high success rate, making the system highly vulnerable to that particular scenario. Conversely, low attack vulnerability implies a low success rate for the evaluated attack scenario, indicating that the system is relatively less vulnerable to that scenario.

2) ATTACK EXECUTION DIFFICULTY

Attack execution difficulty is a metric used to quantitatively assess the challenge of attacking a biometric authentication system. The amount and types of information needed for executing adversarial attacks can vary depending on the attack method. For instance, black-box attacks do not require information about the target system's structure or training parameters, unlike white-box attacks.

This metric is based on the types of information required for executing the attack. It measures whether each type of information is necessary ($x_i \in \{0, 1\}$) and the difficulty of collecting it (w_i). The difficulty of collecting each type of information can be configured based on the user environment, such as the environment in which the target system is deployed and the access control status of each type of information ($\sum_{i=1}^{6} w_i = 100\%$). Through systematic analysis of existing studies [4], [5], [24], [117], we identified six types of information required for attack execution: model architecture, parameters, training data, output result, target label, and training algorithm (see Table 3). The formula is as follows:

$$(Attack\ execution\ difficulty) = \sum_{i=1}^{6} w_i x_i \quad (2)$$

TABLE 4. **Description of defense strategies against adversarial attacks.**

No.	Defense Strategy	Description
1	Adversarial training [117]	A defense strategy in which the model trains a dataset with adversarial examples, to enhance the robustness of the model against adversarial attacks.
2	Defensive distillation [117]	A defense strategy that uses the probability vector generated by a distillation model as the label, to enhance the robustness of the model against adversarial attacks.
3	Gradient regularization [24]	A defense strategy in which the model trains a dataset with slight variations, to enhance the robustness of the model against adversarial attacks.
4	Gradient masking [118]	A defense strategy that masks the model's gradient, to protect it against adversarial attacks.
5	Auxiliary detection model [118]	A defense strategy that uses an auxiliary model that classifies adversarial examples as a filter, to proactively reject adversarial attacks.
6	Image reconstruction [117]	A defense strategy that reconstructs input data using a GAN, to neutralize adversarial attacks.
7	Image denoising [117]	A defense strategy that removes the perturbation within input data using the denoising methods, to neutralize adversarial attacks.
8	Random noising [119]	A defense strategy that adds random noise to input data, to neutralize adversarial attacks.
9	Ensemble learning [118]	A defense strategy that uses multiple classification models, to enhance the robustness of the model against adversarial attacks.
10	Feature squeezing [24]	A defense strategy that transforms the input data into a compact representation, to neutralize adversarial attacks.

High attack execution difficulty suggests that the attack scenario used in the test requires more information, making it challenging to execute the attack and subsequently rendering the system relatively less vulnerable to the scenario. Conversely, low attack execution difficulty implies less required information for the evaluated attack scenario, indicating that the attack is easier to execute and the system is relatively more vulnerable to that scenario.

3) ATTACK DEFENSE AVAILABILITY

Attack defense availability is a metric used to quantitatively assess the extent to which a biometric authentication system can be defended against adversarial attacks. The effectiveness of defense strategies can vary depending on the attack method [4], [120], [121]. For instance, black-box attacks, which are executed without information about the target system, may render certain defense strategies ineffective (e.g., protecting the model's gradient).

This metric is determined by the number of defense strategies available for each type of adversarial attack. We analyzed and classified defense strategies based on existing research on defending against adversarial attacks. We collected relevant papers using the keyword-based literature search process described in Section III-B. Specifically, we gathered a total of 127 papers related to defense strategies against adversarial attacks using search queries that combined 'adversarial attack' with keywords related to defense (i.e., defense and protection). The authors reviewed these papers to exclude those that were less relevant or unsuitable for our objectives. Ultimately, we selected 30 papers [4], [13], [14], [15], [24], [117], [118], [119], [121], [122], [123], [124], [125], [126], [127], [128], [129], [130], [131], [132], [133], [134], [135], [136], [137], [138], [139], [140], [141], [142] focused on defense strategies against adversarial attacks. Through systematic analysis, we identified ten popular defense strategies: adversarial training, defensive distillation, gradient regularization, gradient masking, auxiliary detection model, image reconstruction, image denoising, random noising, ensemble learning, and feature squeezing (see Table 4). The formula for attack defense availability is as follows:

$$(\textit{Attack defense availability}) = \frac{(\textit{The number of available defense strategies})}{(\textit{The total number of defense strategies})} \times 100 \tag{3}$$

High attack defense availability suggests that the system is relatively less vulnerable to the evaluated attack scenario because more defense strategies are available. Conversely, low attack defense availability indicates vulnerability to the evaluated attack scenario due to the scarcity of available defense strategies.

IV. CASE STUDY

To verify the applicability of our proposed approach, we conducted a case study evaluating the risk of adversarial attacks on a face recognition system through scenario-based adversarial attack simulations.

TABLE 5. The structure of training datasets (FaceForensics++ and CASIA-Webface) for a case study.

Dataset	Target Module	# of Classes	# of Images
FaceForensics++	Liveness detector	2	250,000 (Real: 50,000, Fake: 200,000)
CASIA-Webface	Identity matcher	1,000	40,000 (40 by Identity)

A. EXPERIMENTAL SETTING

1) TARGET SYSTEM

We selected the FaceNet architecture, one of the representative biometric authentication systems, as the foundation for the target face recognition system [21], [143]. The datasets utilized for training the target system comprise FaceForensics++ (FF++) [144] and CASIA-Webface [22], detailed in Table 5.

The liveness detector in the target system was trained using the FF++ dataset, which comprises 1,000 real videos and 4,000 fake videos generated through four methods (i.e., Deepfakes, Face2Face, FaceSwap, and NeuralTextures). To train the liveness detector, we sliced the videos into 50 frames [149], [150], [151], [152], [153]. Meanwhile, the identity matcher in the target system was trained using the CASIA-Webface dataset, which contains 494,414 face images representing 10,575 real identities. Due to the uneven distribution of data per identity in CASIA-Webface, we randomly selected 1,000 identities with a minimum of 44 images each for training the identity matcher. Subsequently, we randomly selected 44 images per identity and split these images into training and testing datasets (training: 40 and testing: 4), as CASIA-Webface is also employed as the testing dataset (further details about the testing dataset can be found in Section IV-A3).

The hyperparameters used for training the target system were set as follows: liveness detector (optimizer: Adam, epoch: 6, batch size: 32, and learning rate: 0.001) and identity matcher (optimizer: Adam, epoch: 30, batch size: 32, and learning rate: 0.001). The optimizer, batch size, and learning rate were set to the default values of FaceNet, while the epochs were optimized according to the size of each training dataset. All experiments were conducted on one GPU (NVIDIA GeForce RTX 3090), using Python 3.8.10 and PyTorch 2.0.0+cu117.

2) ADVERSARIAL ATTACK BASED ON ATTACK SCENARIOS

We selected suitable adversarial attack methods for each scenario, as shown in Table 6, based on the following criteria: (1) the method demonstrated state-of-the-art attack performance, and (2) the method's code was publicly available and operated without errors. Consequently, we selected VNI-FGSM [145], Square [146], Witches' Brew [147], GMI [28], and BREP-MI [148]. The simulations were performed using the default settings provided by the authors of each method.

For scenarios involving training phase-logic corruption (i.e., #1, #3, #8, #9, and #12), we modified parts of the target system's algorithm to conduct adversarial attacks. For scenarios targeting the liveness detector and identity matcher (i.e., #1, #8, and #9), we adjusted the training loss in the training algorithm. For scenarios targeting the decision maker (i.e., #3 and #12), we modified the threshold in the inference algorithm.

3) TESTING DATASET

We constructed three testing datasets based on the goals of the attack scenarios, as shown in Table 7. This is because the individual inputting the biometric trait into the system varies depending on the scenario's goal. For example, in scenarios targeting integrity violation, where the attacker maliciously accesses the target system, the attacker inputs the biometric trait. Conversely, in scenarios targeting availability violation, where the user's access to the target system is denied, the benign user-inputs the biometric trait.

Testing dataset#1 is used to evaluate attack scenarios with the goal of integrity violation (i.e., #4, #5, #8, and #11). These scenarios represent situations where the attacker impersonates a specific user to access the target system. To construct testing dataset#1, we randomly selected ten identities from CASIA-Webface to serve as attackers. The images in testing dataset#1 were taken from the 4 images per identity that were not used in the training dataset. Additionally, since these scenarios require impersonation targets, we randomly selected ten other identities and paired each with one of the attacker identities. Therefore, testing dataset#1 consists of 40 images, with four images for each of the ten attackers. In the adversarial attack simulations using testing dataset#1, an attack is considered successful if the attacker's identity is recognized as the impersonation target.

Testing dataset#2 is used to evaluate attack scenarios with the goal of availability violation (i.e., #1, #2, #3, #9, #10, and #12). These scenarios represent situations where the user is unable to access the target system. To construct testing dataset#2, we randomly selected 1,000 identities from CASIA-Webface to serve as users. The images in testing dataset#2 were taken from the 4 images per identity that were not used in the training dataset. Note that, testing dataset#2 consists only of real images. Therefore, testing dataset#2 contains 4,000 images, with four images for each of the 1,000 user identities. In the adversarial attack simulations using testing dataset#2, an attack is considered successful if the user's identity fails to authenticate with the target system.

Testing dataset#3 is used to evaluate attack scenarios with the goal of privacy violation (i.e., #6 and #7). These scenarios represent situations where the target system's training data is extracted based on the attacker's inputs and the system's outputs. To construct testing dataset#3, we utilized images extracted through adversarial attacks. Therefore, testing dataset#3 contains 1,000 images, each corresponding to a unique user identity extracted through adversarial attacks. In the adversarial attack simulations using testing dataset#3,

TABLE 6. Adversarial attack methods for various scenario-based adversarial attack simulations.

Adversarial Attack Method	Goal	Type	Capability		Scenario No.
			Phase	Attack Vector	
VNI-FGSM [145]	Integrity violation	Evasion attack	Testing phase	White-box	#4
Square [146]	Integrity violation	Evasion attack	Testing phase	Black-box	#5
Witches' Brew [147]	Integrity violation / Availability violation	Poisoning attack	Training phase	Data modification and injection	#2, #10, #11
Modification of algorithm	Integrity violation / Availability violation	Poisoning attack	Training phase	Logic corruption	#1, #3, #8, #9, #12
GMI [28]	Privacy violation	Exploratory attack	Testing phase	White-box	#6
BREP-MI [148]	Privacy violation	Exploratory attack	Testing phase	Black-box	#7

TABLE 7. The structure of testing datasets by the goals of the attack scenario for a case study.

No.	Dataset	Type of Goal	# of Classes	# of Images	Scenario No.
1	CASIA-Webface	Integrity violation	10	40 (4 by identity)	#4, #5, #8, #11
2	CASIA-Webface	Availability violation	1,000	4,000 (4 by identity)	#1, #2, #3, #9, #10, #12
3	Extracted image using GMI or BREP-MI	Privacy violation	1,000	1,000 (1 by identity)	#6, #7

TABLE 8. The results of the risk evaluation of adversarial attacks for scenarios with the goal of integrity violation.

Scenario No.	Attack Vulnerability (↑)	Attack Execution Difficulty (↓)	Attack Defense Availability (↓)
#4	**90.00%**	50.01%	100.00%
#5	82.50%	**16.67%**	90.00%
#8	60.00%	33.34%	**0.00%**
#11	0.00%	50.01%	100.00%

an attack is considered successful if the extracted image matches the corresponding identity.

B. EXPERIMENTAL RESULTS

We assessed the risk of adversarial attacks for each attack scenario in the target system using three evaluation metrics: attack vulnerability, attack execution difficulty, and attack defense availability. Note that, for attack execution difficulty, we assumed that the attacker has access to all information in the target system, knowing the location and form of each piece of information, and applied the same difficulty of collecting each type of information (i.e., 16.67%).

1) EVALUATION OF INTEGRITY VIOLATION SCENARIOS

The results of evaluating the risk of adversarial attacks for scenarios with the goal of integrity violation using testing dataset#1 are presented in Table 8. The scenarios include Scenario#4, Scenario#5, Scenario#8, and Scenario#11 (details of each scenario can be found in Section III-C).

For attack vulnerability, Scenario#4 exhibited the highest vulnerability level compared to other scenarios: Scenario#4 (90.00%, 36/40), Scenario#5 (82.50%, 33/40), Scenario#8 (60.00%, 24/40), and Scenario#11 (0.00%, 0/40). Therefore, the target system is most vulnerable to Scenario#4 in terms of attack vulnerability.

For attack execution difficulty, Scenario#5 demonstrated the lowest difficulty level compared to other scenarios: Scenario#4 (50.01%, 3/6), Scenario#5 (16.67%, 1/6), Scenario#8 (33.34%, 2/6), and Scenario#11 (50.01%, 3/6). Therefore, the target system is most vulnerable to Scenario#5 in terms of attack execution difficulty.

For attack defense availability, Scenario#8 showed the lowest availability level compared to other scenarios: Scenario#4 (100.00%, 10/10), Scenario#5 (90.00%, 9/10), Scenario#8 (0.00%, 0/10), and Scenario#11 (100.00%, 10/10). Therefore, the target system is most vulnerable to Scenario#8 in terms of attack defense availability.

Scenario#4 showed higher attack vulnerability compared to Scenario#5. However, Scenario#5 exhibited lower attack execution difficulty and attack defense availability than Scenario#4. Both scenarios entail situations where the attacker inputs an adversarial biometric trait crafted to be recognized as a specific identity. The distinction between the two lies in the information available about the identity matcher (Scenario#4: white-box attack and Scenario#5:

TABLE 9. The results of the risk evaluation of adversarial attacks for scenarios with the goal of availability violation.

Scenario No.	Attack Vulnerability (↑)	Attack Execution Difficulty (↓)	Attack Defense Availability (↓)
#1	82.95%	**16.67%**	**0.00%**
#2	12.68%	33.34%	100.00%
#3	**100.00%**	**16.67%**	**0.00%**
#9	95.43%	**16.67%**	**0.00%**
#10	13.98%	33.34%	100.00%
#12	95.43%	**16.67%**	**0.00%**

TABLE 10. The results of the risk evaluation of adversarial attacks for scenarios with the goal of privacy violation.

Scenario No.	Attack Vulnerability (↑)	Attack Execution Difficulty (↓)	Attack Defense Availability (↓)
#6	16.00%	50.01%	30.00%
#7	**21.70%**	**16.67%**	**20.00%**

black-box attack). Since the white-box attack is executed with the attacker possessing information about the target model, it typically outperforms the black-box attack, where this information is unknown [154]. However, Scenario#4 requires three pieces of information (model architecture, parameters, and target label) for attack execution, while Scenario#5 requires only one (target label). Additionally, ten defense strategies are available against Scenario#4, while nine strategies excluding gradient masking are available against Scenario#5. Despite Scenario#5 having lower attack vulnerability than Scenario#4, it poses a higher risk in terms of attack execution difficulty and attack defense availability. Consequently, the target system is considered more vulnerable to Scenario#5 than to Scenario#4.

Scenario#8 exhibited significantly lower attack defense availability than other scenarios. It depicts a situation where the attacker performs logic corruption by intervening in the training phase of the identity matcher and modifying part of the training algorithm to recognize the attacker's input as a specific identity. To the best of our knowledge, no defense strategy has been proposed for logic corruption at this time. Therefore, the attack defense availability for Scenario#8 is 0.00%.

2) EVALUATION OF AVAILABILITY VIOLATION SCENARIOS

The results of evaluating the risk of adversarial attacks for scenarios with the goal of availability violation using testing dataset#2 are presented in Table 9. The scenarios include Scenario#1, Scenario#2, Scenario#3, Scenario#9, Scenario#10, and Scenario#12 (details of each scenario can be found in Section III-C).

For attack vulnerability, Scenario#3 exhibited the highest vulnerability level compared to other scenarios: Scenario#1 (82.95%, 3,318/4,000), Scenario#2 (12.68%, 507/4,000), Scenario#3 (100.00%, 4,000/4,000), Scenario#9 (95.43%, 3,817/4,000), Scenario#10 (13.98%, 559/4,000), and Scenario#12 (95.43%, 3,817/4,000). Therefore, the target system is most vulnerable to Scenario#3 in terms of attack vulnerability.

For attack execution difficulty, Scenarios#1, #3, #9, and #12 demonstrated the lowest difficulty level compared to other scenarios: Scenario#1 (16.67%, 1/6), Scenario#2 (33.34%, 2/6), Scenario#3 (16.67%, 1/6), Scenario#9 (16.67%, 1/6), Scenario#10 (33.34%, 2/6), and Scenario#12 (16.67%, 1/6). Therefore, the target system is most vulnerable to Scenario#1, #3, #9, and #12 in terms of attack execution difficulty.

For attack defense availability, Scenarios#1, #3, #9, and #12 showed the lowest availability level compared to other scenarios: Scenario#1 (0.00%, 0/10), Scenario#2 (100.00%, 10/10), Scenario#3 (0.00%, 0/10), Scenario#9 (0.00%, 0/10), Scenario#10 (100.00%, 10/10), and Scenario#12 (0.00%, 0/10). Therefore, the target system is most vulnerable to Scenario#1, #3, #9, and #12 in terms of attack defense availability.

Scenarios#1, #3, #9, and #12 exhibited significantly lower attack defense availability than other scenarios. It depicts a situation where the attacker performs logic corruption by intervening in the training phase of the liveness detector (or identity matcher) and modifying part of the training algorithm to prevent users from accessing the system. To the best of our knowledge, no defense strategy has been proposed for logic corruption at this time. Therefore, the attack defense availability for Scenarios#1, #3, #9, and #12 is 0.00%.

TABLE 11. Five risk grades for categorizing the evaluation results from each metric for comprehensive risk analysis of adversarial attacks.

Risk Grade	Attack Vulnerability	Attack Execution Difficulty	Attack Defense Availability
1	$0 \leq$ risk < 20	$80 \leq$ risk ≤ 100	$80 \leq$ risk ≤ 100
2	$20 \leq$ risk < 40	$60 \leq$ risk < 80	$60 \leq$ risk < 80
3	$40 \leq$ risk < 60	$40 \leq$ risk < 60	$40 \leq$ risk < 60
4	$60 \leq$ risk < 80	$20 \leq$ risk < 40	$20 \leq$ risk < 40
5	$80 \leq$ risk ≤ 100	$0 \leq$ risk < 20	$0 \leq$ risk < 20

3) EVALUATION OF PRIVACY VIOLATION SCENARIOS

The results of evaluating the risk of adversarial attacks for scenarios with the goal of privacy violation using testing dataset#3 are presented in Table 10. The scenarios include Scenario#6 and Scenario#7 (details of each scenario can be found in Section III-C).

For attack vulnerability, Scenario#7 exhibited the highest vulnerability level compared to Scenario#6: Scenario#6 (16.00%, 160/1,000) and Scenario#7 (21.70%, 217/1,000). Therefore, the target system is most vulnerable to Scenario#7 in terms of attack vulnerability.

For attack execution difficulty, Scenario#7 demonstrated the lowest difficulty level compared to Scenario#6:

TABLE 12. Comprehensive risk analysis results for a case study, presenting the risk grades of each metric and average risk grades for each scenario.

Goal	Scenario No.	Risk Grade of Attack Vulnerability	Risk Grade of Attack Execution Difficulty	Risk Grade of Attack Defense Availability	Average Risk Grade
Integrity violation	#4	**5 (90.00%)**	3 (50.01%)	1 (100.00%)	3
	#5	**5 (82.50%)**	**5 (16.67%)**	1 (90.00%)	3.7
	#8	4 (60.00%)	4 (33.34%)	**5 (0.00%)**	4.3
	#11	1 (0.00%)	3 (50.01%)	1 (100.00%)	1.7
Availability violation	#1	**5 (82.95%)**	**5 (16.67%)**	**5 (0.00%)**	**5**
	#2	1 (12.68%)	4 (33.34%)	1 (100.00%)	2
	#3	**5 (100.00%)**	**5 (16.67%)**	**5 (0.00%)**	**5**
	#9	**5 (95.43%)**	**5 (16.67%)**	**5 (0.00%)**	**5**
	#10	1 (13.98%)	4 (33.34%)	1 (100.00%)	2
	#12	**5 (95.43%)**	**5 (16.67%)**	**5 (0.00%)**	**5**
Privacy violation	#6	1 (16.00%)	3 (50.01%)	4 (30.00%)	2.7
	#7	2 (21.70%)	**5 (16.67%)**	4 (20.00%)	3.7

Scenario#6 (50.01%, 3/6) and Scenario#7 (16.67%, 1/6). Therefore, the target system is most vulnerable to Scenario#7 in terms of attack execution difficulty.

For attack defense availability, Scenario#7 showed the lowest availability level compared to Scenario#6: Scenario#6 (30.00%, 3/10) and Scenario#7 (20.00%, 2/10). Therefore, the target system is most vulnerable to Scenario#7 in terms of attack defense availability.

Scenario#7 with the black-box attack exhibited higher attack vulnerability than Scenario#6 with the white-box attack. Since the white-box attack is executed with the attacker possessing information about the target model, it typically outperforms the black-box attack, where this information is unknown [154]. However, the BREP-MI used in Scenario#7 outperforms the GMI used in Scenario#6 in terms of attack, resulting in an attack vulnerability contrary to the typical one [148]. Moreover, both attack execution difficulty and attack defense availability are lower in Scenario#7 than in Scenario#6. Specifically, Scenario#6 requires three pieces of information (model architecture, parameters, and target label) for attack execution, while Scenario#7 requires only one (target label). Additionally, three defense strategies are available against Scenario#6, while two strategies are available against Scenario#7. Consequently, the target system is considered more vulnerable to Scenario#7 than to Scenario#6.

4) COMPREHENSIVE ANALYSIS

To comprehensively analyze the risk of adversarial attacks for each scenario, we categorized the evaluation results from each metric into five risk grades, as shown in Table 11 [155], [156]. Table 12 presents the final results of the risk analysis on the target system, showing the average risk grades of each metric for each scenario.

Attack scenarios with an average risk grade of 4 or higher share a common characteristic: the capability for training phase-logic corruption (#1, #3, #8, #9, and #12). These scenarios tend to exhibit relatively higher risk grades across the three metrics for the following reasons. First, the system's logic can be compromised simply by accessing its algorithm, leading to higher risk grades in attack vulnerability and attack execution difficulty. Additionally, no effective defense strategy proposed so far, and identifying the modified parts in the system code is challenging, resulting in a higher risk grade in attack defense availability. To protect the target system from such attacks, it is crucial to manage access to the system to prevent unauthorized entry by attackers [157] and to systematically conduct configuration management to detect changes in files [158].

Attack scenarios with an average risk grade of 2 or lower share a common characteristic: the capability for training phase-data modification and injection (#2, #10, and #11). These scenarios tend to exhibit relatively lower average risk grades for the following reasons. First, detailed information about the target system is required, such as training data, the number of labels, and parameters, while the attack success rate is relatively low, leading to a lower risk grade in attack vulnerability and a medium risk grade in attack execution difficulty. Additionally, various effective defense strategies have been proposed (e.g., preprocessing the training dataset, detecting adversarial examples in the dataset, and protecting parameters of the target model) [4], [120], [121], resulting in a lower risk grade in attack defense availability. To protect the target system from such attacks, it is required to validate the training dataset to prevent data pollution and to employ robust training methods, making poisoning attacks hard to succeed [121].

V. DISCUSSION

In this paper, we defined adversarial attack scenarios based on three criteria: goal, type, and capability. For the goal criterion, various categorizations have been proposed, including: (1) confidence reduction, misclassification, targeted misclassification, and source/target misclassification [4]; and (2) targeted and untargeted attacks [159]. Focusing on existing research on adversarial attacks on biometric authentication systems, we categorized the goals as integrity violation, availability violation, and privacy violation [5].

Among the metrics for assessing the risk of adversarial attacks, attack execution difficulty is defined based on the required information for executing attacks, including model architecture, parameters, training data, output result, target label, and training algorithm. Additionally, attack defense availability depends on the number of available defense strategies against adversarial attacks, such as adversarial training, defensive distillation, gradient regularization, gradient masking, auxiliary detection model, image reconstruction, and image denoising. Our proposed approach can be expanded by incorporating additional required information and defense strategies not previously included. This expansion can be applied when new types of biometric authentication systems are introduced or when new defense strategies emerge.

VI. CONCLUSION

In this paper, we proposed a novel approach for assessing the risk of adversarial attacks on deep learning-based biometric authentication systems. We defined adversarial attack scenarios considering the dependencies between modules of the biometric authentication systems. Our proposed approach evaluates the risk of adversarial attacks from multiple perspectives through adversarial attack simulations based on the defined scenarios. We conducted a case study that assesses the risk of adversarial attacks on a face recognition system implemented with FaceNet. As a result, we confirmed that the attack scenarios with logic corruption exhibit the highest risk of adversarial attacks in terms of vulnerability, execution, and defense.

Our proposed approach enables a systematic security assessment for the target biometric authentication system by evaluating the risk of adversarial attacks based on attack scenarios and evaluation metrics defined in this paper. The results analyzed through our approach enable the establishment of an efficient defense mechanism by identifying the vulnerabilities of adversarial attacks on the target biometric authentication system. Therefore, our approach ultimately enhances the robustness against adversarial attacks.

In our future work, we plan to expand our approach by incorporating additional factors that could influence the risk of adversarial attacks, such as attack execution time and the potential for attack propagation. Furthermore, we aim to develop an automated framework to implement our approach.

REFERENCES

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8647–8695, Aug. 2023.

[2] B. Bhanu and A. Kumar, *Deep Learning for Biometrics*, vol. 7. Springer, 2017.

[3] J. Fei, Z. Xia, P. Yu, and F. Xiao, "Adversarial attacks on fingerprint liveness detection," *EURASIP J. Image Video Process.*, vol. 2020, no. 1, p. 1, Dec. 2020.

[4] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Trans. Intell. Technol.*, vol. 6, no. 1, pp. 25–45, Mar. 2021.

[5] B. Biggio, G. Fumera, P. Russu, L. Didaci, and F. Roli, "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 31–41, Sep. 2015.

[6] S. Marrone, R. Casula, G. Orrù, G. L. Marcialis, and C. Sansone, "Fingerprint adversarial presentation attack in the physical domain," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*. Montréal, CA, USA: Springer, 2021, pp. 530–543.

[7] F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, "Adversarial attacks against face recognition: A comprehensive study," *IEEE Access*, vol. 9, pp. 92735–92756, 2021.

[8] X. Gong, G. Hu, T. Hospedales, and Y. Yang, "Adversarial robustness of open-set recognition: Face recognition and person re-identification," in *Proc. Comput. Vis. Workshops (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 135–151.

[9] T. V. Hamme, G. Garofalo, S. Joos, D. Preuveneers, and W. Joosen, "AI for biometric authentication systems," in *Security and Artificial Intelligence: A Crossdisciplinary Approach*. Springer, 2022, pp. 156–180.

[10] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, "Characterizing and evaluating adversarial examples for offline handwritten signature verification," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2153–2166, Aug. 2019.

[11] S. Joos, T. Van Hamme, D. Preuveneers, and W. Joosen, "Adversarial robustness is not enough: Practical limitations for securing facial authentication," in *Proc. ACM Int. Workshop Secur. Privacy Analytics*, Apr. 2022, pp. 2–12.

[12] A. Musa, K. Vishi, and B. Rexha, "Attack analysis of face recognition authentication systems using fast gradient sign method," *Appl. Artif. Intell.*, vol. 35, no. 15, pp. 1346–1360, Dec. 2021.

[13] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155161–155196, 2021.

[14] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Appl. Sci.*, vol. 9, no. 5, p. 909, Mar. 2019.

[15] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, Apr. 2020.

[16] G. Lovisotto, S. Eberz, and I. Martinovic, "Biometric backdoors: A poisoning attack against unsupervised template updating," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Sep. 2020, pp. 184–197.

[17] A. K. Jain, D. Deb, and J. J. Engelsma, "Biometrics: Trust, but verify," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 3, pp. 303–323, Jul. 2022.

[18] A. Almehmadi, "A behavioral-based fingerprint liveness and willingness detection system," *Appl. Sci.*, vol. 12, no. 22, p. 11460, Nov. 2022.

[19] P. Matthew, "Autonomous synergy with biometric security and liveness detection," in *Proc. Sci. Inf. Conf.*, Oct. 2013, pp. 376–382.

[20] T. Yang, X. Zhao, X. Wang, and H. Lv, "Evaluating facial recognition web services with adversarial and synthetic samples," *Neurocomputing*, vol. 406, pp. 378–385, Sep. 2020.

[21] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[22] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[23] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.

[24] S. Y. Khamaiseh, D. Bagagem, A. Al-Alaj, M. Mancino, and H. W. Alomari, "Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification," *IEEE Access*, vol. 10, pp. 102266–102291, 2022.

[25] D. Deb, V. Mistry, and R. Parthe, "AdvBiom: Adversarial attacks on biometric matchers," 2023, *arXiv:2301.03966*.

[26] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," 2021, *arXiv:2105.03162*.

[27] M. Xue, C. He, J. Wang, and W. Liu, "LOPA: A linear offset based poisoning attack method against adaptive fingerprint authentication system," *Comput. Secur.*, vol. 99, Dec. 2020, Art. no. 102046.

[28] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 250–258.

[29] L. T. P. Lazimul and D. L. Binoy, "Fingerprint liveness detection using convolutional neural network and fingerprint image enhancement," in *Proc. Int. Conf. Energy, Commun., Data Analytics Soft Comput. (ICECDS)*, Aug. 2017, pp. 731–735.

[30] S. M. Lakshmi, M. Kaur, A. K. Shukla, and N. Pathania, "Evasion attack for fingerprint biometric system and countermeasure," in *Proc. Int. Conf. Innov. Comput. Commun.*, Bengaluru, A. Khanna, D. Gupta, S. Bhattacharyya, V. Snasel, J. Platos, and A. E. Hassanien, Eds., Springer, 1007, pp. 71–86.

[31] W. Verheyen, T. Van Hamme, S. Joos, D. Preuveneers, and W. Joosen, "Beware the doppelgänger: Attacks against adaptive thresholds in facial recognition systems," in *Proc. 18th Int. Conf. Availability, Rel. Secur.*, 2023, pp. 1–11.

[32] B. Zi Hao Zhao, H. Jameel Asghar, and M. Ali Kaafar, "On the resilience of biometric authentication systems against random inputs," 2020, *arXiv:2001.04056*.

[33] A. Jardine, T. D. Ramotsoela, and G. P. Hancke, "Biometric authentication system for industrial applications using iris recognition," in *Proc. IEEE 30th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2021, pp. 01–06.

[34] G. C. Fernandez and A. S. Danko, "Addressing the vulnerabilities of pass-thoughts," *Proc. SPIE*, vol. 9842, pp. 507–516, May 2016.

[35] F. K. Carvalho Ota, J. A. Meira, C. R. Cassagnes, and R. State, "Mobile app to SGX enclave secure channel," in *Proc. IEEE Int. Symp. Softw. Rel. Eng. Workshops (ISSREW)*, Oct. 2019, pp. 258–263.

[36] T. Van Hamme, G. Garofalo, D. Preuveneers, and W. Joosen, "Masterkey attacks against free-text keystroke dynamics and security implications of demographic factors," in *Proc. IEEE 8th Eur. Symp. Secur. Privacy (EuroSP)*, Jul. 2023, pp. 278–291.

[37] M. Gofman, S. Mitra, B. Tadesse, and M. Villa, "Biometrics for enterprise security risk mitigation," in *Advances in Cybersecurity Management*. Springer, 2021, pp. 163–195.

[38] S. Zuo, S. Sigg, L. N. Nguyen, N. Beck, N. Jähne-Raden, and M. C. Wolf, "CardioID: Secure ECG-BCG agnostic interaction-free device pairing," *IEEE Access*, vol. 10, pp. 128682–128696, 2022.

[39] E. Lavens, D. Preuveneers, and W. Joosen, "Mitigating undesired interactions between liveness detection components in biometric authentication," in *Proc. 18th Int. Conf. Availability, Rel. Secur.*, Aug. 2023, pp. 1–8.

[40] R. Jain and C. Kant, "Attacks on biometric systems: An overview," *Int. J. Adv. Sci. Res.*, vol. 1, no. 7, p. 283, Sep. 2015.

[41] M. Joshi, B. Mazumdar, and S. Dey, "A comprehensive security analysis of match-in-database fingerprint biometric system," *Pattern Recognit. Lett.*, vol. 138, pp. 247–266, Oct. 2020.

[42] A. Sandirakumaran and A. Kulkarni, "Defending against advanced attack vectors on biometric and authentication systems," in *Proc. AIP Conf.*, vol. 2519, 2022, Paper 030063.

[43] E. Emmanuel, D. Edebatu, N. Catherine, and A. Ngozi, "Vulnerability of biometric authentication system," *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 5, no. 3, pp. 2742–2749, Mar. 2016.

[44] J. C. Bernal-Romero, J. M. Ramirez-Cortes, J. D. J. Rangel-Magdaleno, P. Gomez-Gil, H. Peregrina-Barreto, and I. Cruz-Vega, "A review on protection and cancelable techniques in biometric systems," *IEEE Access*, vol. 11, pp. 8531–8568, 2023.

[45] Y.-H. Baek, B. Kim, and S.-H. Kim, "Fake fingerprint detection biometric system using neural network algorithm," *Int. J. Signal Process. Syst.*, vol. 6, no. 4, pp. 27–30, Dec. 2018.

[46] D. C. Kant, "Liveness detection in different biometric traits: An overview," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, pp. 2011–2015, 2017.

[47] Kavita, G. S. Walia, and R. Rohilla, "A contemporary survey of multimodal presentation attack detection techniques: Challenges and opportunities," *Social Netw. Comput. Sci.*, vol. 2, no. 1, pp. 1–7, Feb. 2021.

[48] S. M. Abdullahi, S. Sun, H. Wang, and B. Wang, "The reversibility of cancelable biometric templates based on iterative perturbation stochastic approximation strategy," *Pattern Recognit. Lett.*, vol. 172, pp. 221–229, Aug. 2023.

[49] Z. Rui and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *IEEE Access*, vol. 7, pp. 5994–6009, 2019.

[50] G. Jagdev and A. Kumar, "Analyzing 2D & 3D fingerprint recognition techniques as secure biometric," *Hand*, vol. 3, no. 3, p. 2.

[51] M. Ghafourian, J. Fierrez, R. Vera-Rodriguez, I. Serna, and A. Morales, "OTB-morph: One-time biometrics via morphing applied to face templates," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 321–329.

[52] K. Shaheed, P. Szczuko, M. Kumar, I. Qureshi, Q. Abbas, and I. Ullah, "Deep learning techniques for biometric security: A systematic review of presentation attack detection systems," *Eng. Appl. Artif. Intell.*, vol. 129, Mar. 2024, Art. no. 107569.

[53] D. F. Smith, A. Wiliem, and B. C. Lovell, "Face recognition on consumer devices: Reflections on replay attacks," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 736–745, Apr. 2015.

[54] K. Sadeghi, A. Banerjee, and S. K. S. Gupta, "A system-driven taxonomy of attacks and defenses in adversarial machine learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 4, pp. 450–467, Aug. 2020.

[55] D. Sharma and A. Selwal, "A survey on face presentation attack detection mechanisms: Hitherto and future perspectives," *Multimedia Syst.*, vol. 29, no. 3, pp. 1527–1577, Jun. 2023.

[56] D. Dasgupta, A. Roy, and A. Nag, *Biometric Authentication*. Cham, Switzerland: Springer, 2017, pp. 37–84.

[57] R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, and G. Sforza, "Biometric recognition in automated border control: A survey," *ACM Comput. Surveys (CSUR)*, vol. 49, no. 2, pp. 1–39, 2016.

[58] S. U. Anayat and A. Selwal, "Template attacks and protection in multi-biometric system: A systematic review," in *Proc. Recent Innov. Comput. (ICRIC)*, 2020, pp. 831–843.

[59] R. Mehmood and A. Selwal, "Fingerprint biometric template security schemes: Attacks and countermeasures," in *Proc. Recent Innov. Comput. (ICRIC)*. Jammu, IN, USA: Springer, 2019, pp. 455–467.

[60] S. Rafiq and A. Selwal, "Template security in iris recognition systems: Research challenges and opportunities," in *Proc. Recent Innov. Comput. (ICRIC)*, 2019, pp. 771–784.

[61] K. Loukhaoukha, A. Refaey, K. Zebbiche, and A. Shami, "Efficient and secure cryptosystem for fingerprint images in wavelet domain," *Multimedia Tools Appl.*, vol. 77, no. 8, pp. 9325–9339, Apr. 2018.

[62] S. Shin and Y. Seto, "Study of cancelable biometrics in security improvement of biometric authentication system," in *Proc. 14th IFIP TC 8 Int. Conf. (CISIM)*. Warsaw, Poland: Springer, 2015, pp. 547–558.

[63] S. Garg and S. H. Mankad, "Voice liveness detection under feature fusion and cross-environment scenario," *Multimedia Tools Appl.*, vol. 79, nos. 37–38, pp. 26951–26967, Oct. 2020.

[64] S. Marrone and C. Sansone, "On the transferability of adversarial perturbation attacks against fingerprint based authentication systems," *Pattern Recognit. Lett.*, vol. 152, pp. 253–259, Dec. 2021.

[65] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? Adversarial attacks on speaker recognition systems," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2021, pp. 694–711.

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

[67] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *Proc. 21st Int. Workshop Mobile Comput. Syst. Appl.*, Mar. 2020, pp. 9–14.

[68] C. Yuan and B. Cui, "Adversarial attack with adaptive gradient variance for deep fake fingerprint detection," in *Proc. IEEE 24th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2022, pp. 1–6.

[69] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, and S. K. Jha, "Directed adversarial attacks on fingerprints using attributions," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–8.

[70] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on ArcFace face ID system," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 819–826.

[71] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7706–7714.

[72] A. J. Bose and P. Aarabi, "Adversarial attacks on face detectors using neural net based constrained optimization," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2018, pp. 1–6.

[73] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial light projection attacks on face recognition systems: A feasibility study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 814–815.

[74] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[75] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang, "Generating adversarial examples by makeup attacks on face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2516–2520.

[76] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1452–1466, 2021.

[77] M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev, and A. Petiushko, "On adversarial patches: Real-world attack on ArcFace-100 face recognition system," in *Proc. Int. Multi-Conf. Eng., Comput. Inf. Sci. (SIBIRCON)*, Oct. 2019, pp. 0391–0396.

[78] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.

[79] F. V. Massoli, F. Carrara, G. Amato, and F. Falchi, "Detection of face recognition adversarial attacks," *Comput. Vis. Image Understand.*, vol. 202, Jan. 2021, Art. no. 103103.

[80] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa, "Detecting and mitigating adversarial perturbations for robust face recognition," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 719–742, Jun. 2019.

[81] L. Yang, Q. Song, and Y. Wu, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *Multimedia Tools Appl.*, vol. 80, pp. 855–875, Sep. 2021.

[82] A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh, "SmartBox: Benchmarking adversarial detection and mitigation algorithms for face recognition," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–7.

[83] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," 2018, *arXiv:1803.04683*.

[84] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1979–1988.

[85] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Are image-agnostic universal adversarial perturbations for face recognition difficult to detect?" in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–7.

[86] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa, "On the robustness of face recognition algorithms against attacks and bias," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 13583–13589.

[87] Z. Yin, K. Uchida, and S. Deng, "Improving adversarial attacks on face recognition using a modified image translation model," in *Proc. 3rd Int. Conf. Image, Video Signal Process.*, Mar. 2021, pp. 26–31.

[88] R. R. Mekala, A. Porter, and M. Lindvall, "Metamorphic filtering of black-box adversarial attacks on multi-network face recognition models," in *Proc. IEEE/ACM 42nd Int. Conf. Softw. Eng. Workshops*, Jun. 2020, pp. 410–417.

[89] K. Fang and J. Yang, "Robust deep facial attribute prediction against adversarial attacks," in *Proc. 7th Int. Conf. Comput. Artif. Intell.*, Apr. 2021, pp. 202–207.

[90] J. Zhang and J. Hou, "Unpaired image-to-image translation network for semantic-based face adversarial examples generation," in *Proc. Great Lakes Symp. VLSI*, Jun. 2021, pp. 449–454.

[91] K. Agrawal and C. Bhatnagar, "BMIM: Generating adversarial attack on face recognition via binary mask," in *Proc. Int. Conf. Intell. Technol. (CONIT)*, Jun. 2021, pp. 1–5.

[92] G. Zhang, H. Jing, X. Wang, C. Zhou, X. He, and D. Ma, "Thwart physical and digital domain's adversarial attack methods on face detection," in *Proc. 2nd Int. Conf. Artif. Intell. Comput. Eng. (ICAICE)*, Nov. 2021, pp. 861–871.

[93] H. Ding, S. He, Y. Wu, Y. Jin, L. Gan, G. Xu, and H. Yang, "An efficient face recognition attack method based on generative adversarial networks and cosine metrics," in *Proc. 5th Asian Conf. Artif. Intell. Technol. (ACAIT)*, Oct. 2021, pp. 570–577.

[94] H. Yuan, Q. Chu, F. Zhu, R. Zhao, B. Liu, and N. Yu, "Efficient open-set adversarial attacks on deep face recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[95] H. Kwon, O. Kwon, H. Yoon, and K.-W. Park, "Face friend-safe adversarial example on face recognition system," in *Proc. 11th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2019, pp. 547–551.

[96] Y. Kim, D. Han, C. Kim, and H.-J. Yoo, "A 0.22–0.89 mW low-power and highly-secure always-on face recognition processor with adversarial attack prevention," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 5, pp. 846–850, May 2020.

[97] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang, "Exploring frequency adversarial attacks for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4093–4102.

[98] I. Singh, S. Momiyama, K. Kakizaki, and T. Araki, "On brightness agnostic adversarial examples against face recognition systems," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2021, pp. 1–5.

[99] C. Sadu and P. K. Das, "A defense method against facial adversarial attacks," in *Proc. IEEE Region 10 Conf. (TENCON)*, Dec. 2021, pp. 459–463.

[100] C. Sadu and P. K. Das, "Detection of geometry-based adversarial facial attacks using error level analysis," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2022, pp. 1–6.

[101] M. Shen, H. Yu, L. Zhu, K. Xu, Q. Li, and J. Hu, "Effective and robust physical-world attacks on deep learning face recognition systems," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4063–4077, 2021.

[102] Y. Liao, L. Huang, and N. Liu, "Evaluating robustness of 3D face reconstruction against adversarial attacks," in *Proc. 9th Int. Conf. Digit. Home (ICDH)*, Oct. 2022, pp. 155–162.

[103] H. Wang, S. Wang, Z. Jin, Y. Wang, C. Chen, and M. Tistarelli, "Similarity-based gray-box adversarial attack against deep face recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.

[104] X. Wang, R. Ni, W. Li, and Y. Zhao, "Adversarial attack on fake-faces detectors under white and black box scenarios," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3627–3631.

[105] J. Byun, H. Go, and C. Kim, "Geometrically adaptive dictionary attack on face recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Sep. 2022, pp. 3021–3030.

[106] H. Mav, A. Mokashi, S. Nanduri, and V. Pinjarkar, "Face recognition and adversarial masking techniques," in *Proc. 3rd Int. Conf. Emerg. Technol. (INCET)*, May 2022, pp. 1–7.

[107] F. Ding, B. Fan, Z. Shen, K. Yu, G. Srivastava, K. Dev, and S. Wan, "Securing facial bioinformation by eliminating adversarial perturbations," *IEEE Trans. Ind. Informat.*, vol. 19, no. 5, pp. 6682–6691, May 2023.

[108] C. Zhou, H. Jing, X. He, L. Wang, K. Chen, and D. Ma, "Disappeared face: A physical adversarial attack method on black-box face detection models," in *Proc. 23rd Int. Conf. Inf. Commun. Security (ICICS)*. Chongqing, China: Springer, 2021, pp. 119–135.

[109] Z. Chen, P. Lin, Z. L. Jiang, Z. Wei, S. Yuan, and J. Fang, "An illumination modulation-based adversarial attack against automated face recognition system," in *Proc. 16th Int. Conf. Inf. Secur. Cryptol., Inscrypt*. Guangzhou, China: Springer, 2021, pp. 53–69.

[110] S. A. Kilany, A. Mahfouz, A. M. Zaki, and A. Sayed, "Analysis of adversarial attacks on face verification systems," in *Proc. Int. Conf. Artif. Intell. Comput. Vis.* Settat, MA, USA: Springer, 2021, pp. 463–472.

[111] E. Lyko and M. Kedziora, "Adversarial attacks on face detection algorithms using anti-facial recognition T-shirts," in *Proc. 13th Int. Conf. (ICCCI)*. Kallithea, Rhodes: Springer, 2021, pp. 266–277.

[112] L. Kurnianggoro and K.-H. Jo, "Ensemble of predictions from augmented input as adversarial defense for face verification system," in *Proc. 11th Asian Conf. (ACIIDS)*. Yogyakarta, Indonesia: Springer, 2019, pp. 658–669.

[113] J. Zhou, C. Liang, and J. Chen, "Manifold projection for adversarial defense on face recognition," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 288–305.

[114] B. Yan, Q. Wu, and Y. Wang, "Disentanglement of deep features for adversarial face detection," in *Proc. Chin. Conf. Biometric Recognit.* Beijing, China: Springer, 2022, pp. 149–157.

[115] X. Yang, F. Wei, H. Zhang, and J. Zhu, "Design and interpretation of universal adversarial patches in face detection," in *Proc. 16th Eur. Conf. Comput. Vision (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 174–191.

[116] D. C. Nguyen, N. D. Le, T. C. Nguyen, T. Q. Nguyen, and V. Q. Nguyen, "An approach to evaluate the reliability of the face recognition process using adversarial samples generated by deep neural networks," in *Proc. Intell. Syst. Netw. Sel. Articles (ICISN)*. Hanoi, Vietnam: Springer, 2022, pp. 237–245.

[117] H. Liang, E. He, Y. Zhao, Z. Jia, and H. Li, "Adversarial attack and defense: A survey," *Electronics*, vol. 11, no. 8, p. 1283, Apr. 2022.

[118] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey toward the defender's perspective," *ACM Comput. Surveys*, vol. 55, no. 1, pp. 1–38, Nov. 2021.

[119] C. Wang, J. Wang, and Q. Lin, "Adversarial attacks and defenses in deep learning: A survey," in *Proc. Intell. Comput. Theories Appl.*, D.-S. Huang, K.-H. Jo, J. Li, V. Gribova, and V. Bevilacqua, Eds., Cham, Switzerland: Springer, 2021, pp. 450–461.

[120] J. Geiping, L. Fowl, G. Somepalli, M. Goldblum, M. Moeller, and T. Goldstein, "What doesn't kill you makes you Robust(er): How to adversarially train against data poisoning," 2021, *arXiv:2102.13624*.

[121] Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, and K. Ren, "Threats to training: A survey of poisoning attacks and defenses on machine learning systems," *ACM Comput. Surveys*, vol. 55, no. 7, pp. 1–36, Dec. 2022.

[122] X. Wei, B. Pu, J. Lu, and B. Wu, "Visually adversarial attacks and defenses in the physical world: A survey," 2022, *arXiv:2211.01671*.

[123] J. Li, Y. Liu, T. Chen, Z. Xiao, Z. Li, and J. Wang, "Adversarial attacks and defenses on cyber–physical systems: A survey," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5103–5115, Jun. 2020.

[124] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, and Q. Yu, "A survey of adversarial attack and defense methods for malware classification in cyber security," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 467–496, 1st Quart., 2023.

[125] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, and H. Vincent Poor, "Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2245–2298, 1st Quart., 2023.

[126] W. Tan, J. Zhao, X. Liang, H. Lu, B. Song, and H. Guan, "Adversarial example attack and defence of object recognition: A survey," in *Proc. IEEE Int. Conf. Unmanned Syst. (ICUS)*, Oct. 2022, pp. 1–6.

[127] M. Girdhar, J. Hong, and J. Moore, "Cybersecurity of autonomous vehicles: A systematic literature review of adversarial attacks and defense models," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 417–437, 2023.

[128] H. Tan, L. Wang, H. Zhang, J. Zhang, M. Shafiq, and Z. Gu, "Adversarial attack and defense strategies of speaker recognition systems: A survey," *Electronics*, vol. 11, no. 14, p. 2183, Jul. 2022.

[129] A. Alotaibi and M. A. Rassam, "Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense," *Future Internet*, vol. 15, no. 2, p. 62, Jan. 2023.

[130] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Comput. Sci. Rev.*, vol. 37, Aug. 2020, Art. no. 100270.

[131] Y. Li, M. Cheng, C.-J. Hsieh, and T. C. M. Lee, "A review of adversarial attack and defense for classification methods," *Amer. Statistician*, vol. 76, no. 4, pp. 329–345, Oct. 2022.

[132] G. W. Muoka, D. Yi, C. C. Ukwuoma, A. Mutale, C. J. Ejiyi, A. K. Mzee, E. S. A. Gyarteng, A. Alqahtani, and M. A. Al-antari, "A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense," *Mathematics*, vol. 11, no. 20, p. 4272, Oct. 2023.

[133] K. T. Y. Mahima, M. Ayoob, and G. Poravi, "Adversarial attacks and defense technologies on autonomous vehicles: A review," *Appl. Comput. Syst.*, vol. 26, no. 2, pp. 96–106, Dec. 2021.

[134] A. Bajaj and D. K. Vishwakarma, "A state-of-the-art review on adversarial machine learning in image classification," *Multimedia Tools Appl.*, vol. 83, no. 3, pp. 9351–9416, Jan. 2024.

[135] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for DNN models: A review and experimental comparison," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4403–4462, Aug. 2022.

[136] A. Amirkhani, M. P. Karimi, and A. Banitalebi-Dehkordi, "A survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles," *Vis. Comput.*, vol. 39, no. 11, pp. 5293–5307, Nov. 2023.

[137] C. Meyers, T. Löfstedt, and E. Elmroth, "Safety-critical computer vision: An empirical survey of adversarial evasion attacks and defenses on computer vision systems," *Artif. Intell. Rev.*, vol. 56, no. S1, pp. 217–251, Oct. 2023.

[138] A. Michel, S. K. Jha, and R. Ewetz, "A survey on the vulnerability of deep neural networks against adversarial attacks," *Prog. Artif. Intell.*, vol. 11, no. 2, pp. 131–141, Jun. 2022.

[139] M. Mbow, K. Sakurai, and H. Koide, "Advances in adversarial attacks and defenses in intrusion detection system: A survey," in *Proc. Int. Conf. Sci. Cyber Secur.* Matsue, Japan: Springer, 2022, pp. 196–212.

[140] J. Dong, X. Gong, and M. Xue, "Adversarial examples in wireless networks: A comprehensive survey," in *Proc. Int. Conf. Edge Comput. IoT*. Shenzhen, China: Springer, 2021, pp. 92–97.

[141] Z. Feng, C. Liu, X. Ji, and X. Liu, "A survey of adversarial examples and deep learning based data hiding," in *Proc. Int. Symp. Secur. Privacy Social Netw. Big Data*. Fuzhou, China: Springer, 2021, pp. 161–171.

[142] D. Wang, R. Wang, L. Dong, D. Yan, X. Zhang, and Y. Gong, "Adversarial examples attack and countermeasure for speech recognition system: A survey," in *Proc. Int. Conf. Secur. Privacy Digit. Economy*. Quzhou, China: Springer, 2020, pp. 443–468.

[143] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021.

[144] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[145] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1924–1933.

[146] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 484–501.

[147] J. Geiping, L. H. Fowl, W. R. Huang, W. Czaja, G. Taylor, M. Moeller, and T. Goldstein, "Witches' brew: Industrial scale data poisoning via gradient matching," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–24.

[148] M. Kahla, S. Chen, H. A. Just, and R. Jia, "Label-only model inversion attacks via boundary repulsion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15025–15033.

[149] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid transformer network for deepfake detection," in *Proc. 19th Int. Conf. Content-Based Multimedia Indexing*. New York, NY, USA: Association for Computing Machinery, 2022, pp. 8–14.

[150] Z. Guo, G. Yang, J. Chen, and X. Sun, "Exposing deepfake face forgeries with guided residuals," *IEEE Trans. Multimedia*, vol. 25, pp. 8458–8470, 2023.

[151] H. Lin, W. Huang, W. Luo, and W. Lu, "DeepFake detection with multi-scale convolution and vision transformer," *Digit. Signal Process.*, vol. 134, Apr. 2023, Art. no. 103895.

[152] B. Wang, X. Wu, Y. Tang, Y. Ma, Z. Shan, and F. Wei, "Frequency domain filtered residual network for deepfake detection," *Mathematics*, vol. 11, no. 4, p. 816, Feb. 2023.

[153] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, "Exploiting fine-grained face forgery clues via progressive enhancement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 735–743.

[154] Y. Wang, J. Liu, X. Chang, R. J. Rodríguez, and J. Wang, "DI-AA: An interpretable white-box attack for fooling deep neural networks," *Inf. Sci.*, vol. 610, pp. 14–32, Sep. 2022.

[155] A. Gonçalves, M. C. Marques, S. Loureiro, R. Nieto, and M. L. R. Liberato, "Disruption risk analysis of the overhead power lines in Portugal," *Energy*, vol. 263, Jan. 2023, Art. no. 125583.

[156] D. Řehák, P. Danihelka, and A. Bernatik, "Criteria risk analysis of facilities for electricity generation and transmission," in *Proc. Safety, Rel. Risk Anal., Beyond Horizon (ESREL)*, 2014, pp. 2073–2080.

[157] O. A. Farayola, O. L. Olorunfemi, and P. O. Shoetan, "Data privacy and security in it: A review of techniques and challenges," *Comput. Sci. IT Res. J.*, vol. 5, pp. 606–615, Mar. 2024.

[158] J. Larumbe, J. García-Barruetabeña, and D. López-de Ipiña, "The importance of configuration management on complex industrial environments," *New Technol.*, vol. 7, no. 1, p. 14, Jan./Dec. 2020.

[159] P. Rathore, A. Basak, S. H. Nistala, and V. Runkana, "Untargeted, targeted and universal adversarial attacks and defenses on time series," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

SEONG HEE PARK received the B.S. degree in information security from Seoul Women's University, Seoul, Republic of Korea, and the M.S. degree in computer engineering from Hongik University, Seoul, Republic of Korea. Her research interests include deep learning algorithms and their applications to information security.

SOO-HYUN LEE received the B.S. degree in information security from Seoul Women's University, Seoul, Republic of Korea, and the M.S. degree in computer engineering from Hongik University, Seoul. Her research interests include deep learning algorithms and information security.

MIN YOUNG LIM received the B.S. degree in information security from Seoul Women's University, Seoul, Republic of Korea, and the M.S. degree in computer engineering from Hongik University, Seoul. Her research interests include deep learning algorithms and their applications to AI security. She was a recipient of the IEEE ICTC Best Paper Award, in 2022.

PYO MIN HONG received the B.S. degree in information security from Seoul Women's University, Seoul, Republic of Korea. She is currently pursuing the M.S. degree in computer engineering with Hongik University, Seoul. Her research interests include deep learning, computer vision, and information security.

YOUN KYU LEE received the B.S. and M.S. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2010 and 2012, respectively, and the Ph.D. degree in computer science from the University of Southern California (USC), Los Angeles, CA, USA, in 2017. He is currently an Assistant Professor with the Department of Computer Engineering, Hongik University, Seoul, South Korea. Before joining Hongik University, he was with the Samsung Advanced Institute of Technology, Suwon, South Korea, from 2018 to 2020, and Seoul Women's University, Seoul, from 2020 to 2021. He was a recipient of the Viterbi Graduate Fellowship with his Ph.D. admission from USC, in 2012, the IEEE/ACM International Conference on Automated Software Engineering (ASE) Best Tool Paper Award, in 2018, and the IEEE ICTC Best Paper Award, in 2022.