International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

RobustID: A Quantitative Framework for Evaluating the Resilience of Deep Learning Models Against Identity Manipulation

Suman Kumar Sanjeev Prasanna*[1], Shardul Pandya[2]

Submitted:13/08/2022 **Revised**: 20/09/2022 **Accepted**: 29/09/2022

Abstract: As deep learning models increasingly govern operational identity verification, their vulnerability to sophisticated adversarial manipulation presents a critical risk to digital integrity. This research introduces RobustID, a comprehensive evaluation framework designed to quantify and enhance the resilience of neural identity detectors. The framework systematically applies a multi-vector attack taxonomy, including adversarial perturbations (PGD/FGSM), presentation attacks (PA), and cross-modal latent injections across diverse biometric and behavioral datasets. A core technical contribution of RobustID is the integration of Bayesian uncertainty estimation to quantify detection degradation and identify the breaking points of state-of-the-art verification architectures. Furthermore, the study evaluates strategic mitigation regimens, including adversarial training, feature-space regularization, and multimodal redundancy. Empirical results reveal that while standard models are highly susceptible to texture-based and resolution-aware spoofing, the implementation of the RobustID-derived adaptive defenses can substantially improve model robustness without compromising baseline accuracy. By bridging the gap between theoretical adversarial AI and practical security, this research establishes a definitive methodology for deploying resilient, confidence-aware deep learning systems in high-stakes operational environments.

Keywords: *Adversarial Attack, Deep Learning, Identity Recognition, Robustness Evaluation, Security, Temporal Consistency, Transfer Robustness.*

1. Introduction

The tremendous growth in the development and application of deep learning has greatly changed the landscape of identity recognition systems, including biometric verification, surveillance, digital authentication, and financial security [1]. The use of convolutional neural networks and metric learning has greatly allowed for highly accurate facial recognition, speaker verification, and multimodal identity recognition by learning discriminative representations from large data sets [2]. These systems essentially rely on the learning and recognition of high-dimensional feature embeddings that uniquely represent an individual's biometric or behavioral traits. However, with the growing use and deployment of these systems in critical infrastructures and consumer electronics, identity recognition systems often operate in unconstrained scenarios that include changes in lighting, pose, occlusion, aging, compression, and sensor noise [3]. Despite the systems performing almost impeccably under standard testing scenarios, research in artificial intelligence security has shown that deep learning models, despite their prediction capabilities, often have fragile decision boundaries that can be easily affected by a slight and often negligible change in the input data, which raises critical security and reliability issues [4].

[1*,2]*School of Computer and Information Sciences*
University of the Cumberlands
Williamsburg, KY
** Corresponding Author Email: sprasanna68498@ucumberlands.edu*

In parallel, developments in machine learning have led to new paradigms for decision-making under uncertainty. Sequential decision theories such as Markov Decision Processes offer a formal mathematical structure for modeling environments where outcomes are a function of states as well as decisions taken within those states [5]. Further, reinforcement learning extends this paradigm to learn optimal decisions through experience gained through interacting with the environment. However, in adversarial environments, attackers actively seek to subvert the environment by manipulating inputs, states, or transitions to impair performance or evade detection [6]. Identity verification systems are increasingly operating within such environments where adversaries adjust their behavior based on deployed countermeasures. Static authentication pipelines are unable to effectively counter such threats as a result of their inability to learn from experience [7]. Theories of adversarial machine learning reveal how models can be misled through strategic perturbations, as well as theories of security that emphasize the need for risk-adaptive verification systems. The aforementioned developments provide a conceptual basis for understanding identity verification as a sequential adversarial decision-making paradigm where effectiveness is a function of adaptive decisions under uncertainty and shifting adversaries [8].

The purpose and goal of this study are to evaluate the robustness of deep learning-based identity recognition systems with regard to identity manipulation attacks,

specifically focusing on the transfer, temporal, and adversarial perturbation strategies. This study aims to address the following purpose and goal because of the increased usage of identity recognition systems, which can be used for security-critical applications, and the slightest weakness can cause serious problems, including impersonation, privacy violation, and unauthorized access. Despite the high accuracy achieved by the identity recognition systems, there has been minimal attention paid to the robustness benchmarking with regard to various attack strategies. The purpose and goal of the paper are as follows: The main contribution of the paper includes the development of a robust evaluation framework, the evaluation of the transferability, the evaluation of the temporal stability, and the evaluation of the robustness of the identity recognition systems. The rest of the paper will be discussed as follows: Next, the paper will be discussing the related work, followed by the threat model and methodology, the experimental evaluation, the results discussion, and finally the conclusion.

2. Literature Review

The existing body of research on identity manipulation and robustness of deep learning-based identity recognition systems indicates a dynamic research environment wherein the susceptibility of state-of-the-art systems to manipulative attempts has become a focal point of concern. As deep learning techniques continue to dominate the performance of various identity recognition systems, including face recognition systems, the threat of adversarial attacks and manipulations is a major threat to the integrity of these systems. The studies reviewed in this section discuss various aspects of the susceptibility of deep learning-based identity recognition systems to manipulative attempts. Overall, the studies indicate that deep learning models are not as robust as perceived and must be subjected to evaluation for their ability to withstand identity manipulation threats. The following paragraphs introduce critical studies that offer a foundational understanding of identity recognition threats, as applicable to this study's focus on the evaluation of deep learning models for susceptibility to identity manipulation threats [9].

The research paper by Yaoyao Zhong et al. [10] aims at exploring the transferability of adversarial attacks on deep learning-based face recognition systems. The research paper has focused on exploring the feature-level attack strategies and has proposed the DFANet approach, which aims at enhancing the transferability of adversarial attacks. The research has found that deep learning-based face recognition systems are vulnerable to attacks even if the attacker is not aware of the model parameters. Gaurav Goswami et al. [11] have focused on the robustness of the deep learning-based face recognition systems against different types of adversarial attacks. The research has analyzed the robustness of the deep learning-based face recognition systems against different types of attacks, including image perturbations and adversarial attacks. The research has proposed different mechanisms through which adversarial attacks can be detected and mitigated by analyzing the different activations within the network. The research has found that the deep learning-based identity recognition systems are vulnerable even if the attacks are subtle.

Akshay Agarwal, et al. [12] attempt to address the problem of digital identity manipulation attacks, including face swap and face morphing attacks, by applying a novel feature descriptor to detect attacks. This research extends the scope of adversarial attacks to include more realistic attacks, including digital manipulation attacks, which can be easily created using consumer devices, and evaluates the detection effectiveness of adversarial attacks on manipulated face datasets. Vakhshiteh, et al., [13] aim to study adversarial attacks in a real-world context by creating imperceptible perturbations to mimic makeup effects on specific regions of faces to mislead deep learning models in face recognition attacks. This study shows the effectiveness of adversarial attacks in face recognition, including the use of makeup, which can greatly increase the success of attacks in black-box face recognition attacks, indicating the practicality of adversarial identity manipulation attacks, including physical attacks.

Richa Singh et al. [14], which provides a review of various threat classes to face recognition systems, including physical presentation attacks, disguise/make-up, digital adversarial attacks, and morphing attacks, among others. The study provides a discussion on how the various attacks impact face recognition systems, thereby advocating for the consideration of robustness factors in the evaluation of deep learning-based identity recognition systems. In the study by Yuetian Wang et al. [15], the authors develop a system for adversarial attacks, whereby the study demonstrates the creation of digital adversarial faces and glasses to mislead face recognition systems, thereby proving that deep learning-based face recognition systems can be misled to produce incorrect identity results. Therefore, the study shows the practical implications of deep learning-based face recognition systems, thereby advocating for the consideration of robustness in the evaluation of deep learning-based identity recognition systems.

Massoli et al. [16] focuses on the detection of face recognition adversarial attacks by analyzing how adversarial attacks affect face recognition systems based on deep learning techniques and proposing detection techniques for adversarial attacks on face recognition systems. The authors emphasize that adversarial attacks based on feature space attacks are critical threats to face recognition systems and propose detection techniques that can detect adversarial attacks without retraining face

recognition models. Lu Yang et al. [17] focuses on face recognition adversarial attacks based on generative models and proposes an adversarial attack generative model based on attentional generative adversarial networks to create adversarial face images that can deceive face recognition systems and lead to impersonating attacks by attackers on face recognition systems. The authors propose an adversarial attack generative model based on face recognition systems and demonstrate the effectiveness of the proposed model by generating adversarial face images that can deceive face recognition systems.

Fabrizio Falchi et al [18] examine the performance of deep learning-based systems in the context of face recognition and the significant deterioration in performance under adversarial conditions at different image resolutions. While the research is not specifically focused on the issue of model robustness, it is related in the context of the interplay between image quality and adversarial noise. The research is useful in the context of understanding the overall impact of manipulated input on model performance. Pautov, et al. [19] have made an important contribution in the form of an overall book chapter on the topic of Adversarial Attacks on Face Recognition Systems, which presents an overall view of different adversarial attacks on deep learning-based systems in the context of face recognition. The research presents an overview of different attack taxonomies, including perturbation-based attacks, impersonating attacks, and evasion attacks, and the overall impact on state-of-the-art systems.

Table 1. Comparative Studies on Identity Manipulation and Robustness

Study	Methods	Key Findings
[20]	Proposed a deep Siamese network to detect morphing attacks by comparing feature representations between paired face images; applies contrastive learning and distance metrics for classification.	Demonstrated that differential morph attack detection with Siamese features improves detection accuracy over classical classifiers, highlighting vulnerability of face recognition systems to morphing attacks.
[21]	Introduced a decision-based black-box adversarial attack algorithm that crafts perturbations using only hard-label outputs from the target face recognition system.	Showed that evolutionary search attacks can successfully fool state-of-the-art face models without access to gradients, exposing real-world vulnerabilities.
[22]	Developed feature-space and classifier-level adversarial detection networks that distinguish adversarial inputs from genuine face images.	Found that adversarial feature attacks pose serious threats and that detection models can effectively identify malicious inputs, improving robustness.
[23]	Crafted physical-world adversarial stickers using a sticker generator and transformer to attack face recognition models robustly under physical conditions.	Demonstrated that physical adversarial stickers significantly increase the success rate of both dodging and impersonation attacks, revealing real-world attack feasibility.
[24]	Proposed automated face demorphing techniques to reverse morphing attacks and recover original identities from morphed biometric images.	Showed that demorphing improves accuracy of detecting manipulated identities and helps restore integrity in identity verification systems.
[25]	Used deep convolutional networks to extract high-level features for detecting subtle morph artifacts in face images.	Revealed that deep feature extraction significantly improves detection of morphed images compared to conventional texture-based methods, enhancing biometric robustness.

The research gap arises because, while the development and advancement of deep learning-based identity recognition systems have mainly concentrated on increasing the precision and accuracy of the systems under standard benchmark conditions, relatively less attention has been given to evaluating robustness with regard to sophisticated identity manipulation attacks. Some existing research has examined specific types of attacks, like adversarial attacks, morphing, and disguise, in isolation, while few have examined the combined effects of these attacks from different threat models. Further, some research has concentrated more on the mechanisms for generating or detecting attacks, while few have proposed a unified robustness evaluation framework that considers transferability, consistency, and degradation patterns under controlled conditions. The use of different datasets, architectures, and evaluation metrics in existing research also contributes to the research gap, as it is difficult to

compare and generalize the results. Further, since identity recognition systems operate in a dynamic environment with sequential data and black-box deployment scenarios, there has been relatively less research that has examined the persistence of vulnerabilities over time and across unknown models.

3. Methodology

In this section, a specific methodological framework is outlined, aimed at ensuring the evaluation of robustness in identity recognition models, specifically those built upon deep learning, in terms of identity manipulation attacks. The study follows a specific systematic experimental pipeline, starting from data preparation and supervised learning of various models, followed by adversarial generation, transfer analysis, and temporal robustness evaluation. The study has been designed to include robustness-aware training, specifically to evaluate improvements in terms of adversarial exposure. Each of the components of the methodology has been carefully aligned to ensure the evaluation of specific parameters, including performance degradation, transfer success, and robustness in terms of hyperparameters, to ensure a uniform evaluation of all models, thereby extending the robustness evaluation beyond the scope of accuracy measurement in terms of vulnerability patterns and robustness characteristics of the models built upon deep learning architectures.

3.1. Dataset Modeling and Pre-Processing Strategy

The research has designed a structured dataset configuration for structured robustness evaluation. The research combines large-scale face recognition datasets, including VGGFace2, CASIA-WebFace, and LFW, for diversity in terms of pose, lighting, and identity. The research has divided the data into training, validation, and robustness evaluation sets for avoiding information leakage during adversarial robustness evaluation. The identity information is represented as a numerical value, and the images have been resized and normalized for stable gradient behavior.

Equation 1: Input Normalization

$$x' = \frac{x-\mu}{\sigma} \quad (1)$$

This equation scales pixel intensity values by subtracting the dataset mean μ and dividing by the standard deviation σ. The study applies this transformation to reduce distribution variance and improve convergence stability.

For dataset balance, identity frequency normalization is applied.

Equation 2: Class Weight Adjustment

$$w_i = \frac{1}{n_i} \quad (2)$$

Here, n_i represents the number of samples in class i. This equation ensures that underrepresented identities receive higher training importance.

Data augmentation introduces controlled perturbations.

Equation 3: Augmented Sample Generation

$$x_a = x + \delta \quad (3)$$

The term δ denotes small geometric or photometric variations. This step simulates natural variability before adversarial stress testing. Training is conducted using stochastic mini-batches to ensure stable feature learning. The dataset preparation ensures the present work evaluates robustness across realistic and balanced identity conditions.

3.2. Deep Feature Representation Learning

This research uses convolutional neural networks to learn discriminative identity embeddings. The study is based on feature-space separability as a premise for the assessment of robustness. The training of the networks is based on minimizing classification loss as well as compact intra-class distributions.

Equation 1: Softmax Classification

$$P(y = i \mid x) = \frac{e^{zi}}{\sum_j e^{zi}} \quad (4)$$

This equation converts logits ^{zi} into probability values for identity prediction. It guides supervised identity learning.

Equation 2: Cross-Entropy Loss

$$L_{ce} = -logP(y) \quad (5)$$

This loss penalizes incorrect predictions and drives gradient-based optimization during training.

The work also enforces embedding compactness.

Equation 3: Feature Distance

$$D = \| f(x_1) - f(x_2) \| \quad (6)$$

This Euclidean distance measures similarity between two identity embeddings. Smaller distances represent identity consistency.

The study trains models until convergence using adaptive gradient descent. The theoretical foundation relies on representation learning, where separable embeddings are expected to improve recognition but may remain vulnerable near decision boundaries. This forms the baseline for robustness evaluation.

3.3. Adversarial Perturbation Modeling

The research aims to simulate the manipulation of identity through adversarial noise. The current research aims to test the gradient-based attacks by manipulating the pixel values while maintaining the appearance.

Equation 1: Fast Gradient Perturbation

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x L) \quad (7)$$

Here, ϵ controls perturbation magnitude. The gradient direction maximizes classification loss to mislead identity prediction.

Equation 2: Perturbation Constraint

$$\| \delta \| \infty \leq \epsilon \quad (8)$$

This constraint ensures imperceptible noise by bounding the maximum pixel deviation.

The theoretical reasoning states that adversarial examples exploit linearity in high-dimensional feature space. The research studies how small perturbations shift embeddings across decision boundaries.

Equation 3: Targeted Attack Objective

$$L_{target} = -logP(y_t) \quad (9)$$

This loss forces the model to predict a target identity y_t. Training remains unchanged, but robustness testing evaluates prediction degradation under attack influence.

3.4. Transferability and Cross-Model Analysis

In this study, the effectiveness of adversarial examples produced using a particular deep learning structure to fool another independently trained structure is investigated. For this, multiple identity recognition models were trained independently using the same dataset but with different structural configurations. For this study, the adversarial examples were first produced using a source model under a white-box regime. These adversarial examples were then used to test the target model under a black-box regime. This represents a practical scenario because attackers are often unaware of the inner workings of the attacked system. In this study, the generalization of perturbations across heterogeneous models rather than the degradation of a particular model under adversarial attacks was considered. For this, the cross-model tests were conducted without the target network being fine-tuned.

Equation 1: Transfer Success Rate

$$TSR = \frac{Nsuccess}{Ntotal} \quad (10)$$

This equation computes the ratio of adversarial samples that successfully mislead the target model to the total number of transferred samples. A higher TSR indicates stronger transferability and weaker robustness.

Equation 2: Prediction Shift

$$\Delta = | P_{clean} - P_{adv} | \quad (11)$$

This equation measures the absolute change in predicted probability between clean and adversarial inputs. Large shifts indicate instability in the learned decision function.

The transferability of the independently trained networks is based on the fact that they tend to share common feature representations and decision boundaries. The embeddings of the identities exist in high-dimensional feature spaces. The adversarial attacks, which attempt to move an embedding across one decision boundary, can potentially cross similar boundaries in another network. The research quantifies robustness by assessing the cross-architecture vulnerability pattern and embedding sensitivity. Instead of relying on the reduction of accuracy, the research aims to examine the transfer success rate and probability deviation as robustness measures. The research aims to provide an understanding of whether the adversarial threat is model-specific or universally present across the set of identity recognition networks.

3.5. Temporal Robustness Modeling

This research extends the concept of robustness evaluation from the context of static image-based identity recognition to the context of sequential identity systems, which process video streams or frame sequences. Rather than evaluating the adversarial vulnerability of the identity systems, this study focuses on the evaluation of the consistency of predictions over a series of time steps. Temporal modeling captures the practical context of surveillance and identity authentication systems, which often involve the verification of identity based on aggregated evidence. In this study, the concept of gradual adversarial perturbation accumulation over a series of frames is introduced to determine whether the adversarial effect dissipates over time. Each frame of the video stream receives a consistent or increasing level of perturbation.

Equation 1: Temporal Average Prediction

$$P_{seq} = \frac{1}{T} \sum_{t=1}^{T} P_t \quad (12)$$

This equation computes the average identity probability across all frames in a sequence of length TTT. It reflects aggregated decision confidence over time.

Equation 2: Cross-Entropy Loss

$$TSI = 1 - Var(P_t) \quad (13)$$

This equation measures prediction stability by subtracting the variance of frame-level probabilities from one. Lower variance results in higher TSI, indicating stable identity recognition.

The theoretical foundation is based on the assumption that good identity systems should demonstrate stable predictions even with small temporal changes. If there is an effect of adversarial perturbations on the predictions' oscillation and progressive drift, this demonstrates temporal fragility. The current research assesses the issue of identity drift by measuring the changes in class probability predictions with increasing perturbation over time. Temporal robustness modeling introduces a new way of demonstrating vulnerability patterns that would not be captured by static evaluation models. This helps to assess whether adversarial

misclassification is an ongoing, intermittent, or self-correcting phenomenon within sequential decision models.

3.6. Robustness Regularization During Training

This study utilizes robustness-aware optimization for the evaluation of whether training with adversarial exposure improves the model's resistance to identity manipulation. This study integrates adversarial examples into the training process by generating adversarial samples for each mini-batch. Clean and adversarial samples are then used alternatively for balancing the model's recognition capability and robustness improvement. This helps the model learn decision boundaries that are smoother, reducing the sharpness of the decision boundaries with regard to small perturbations.

Equation 1: Adversarial Training Loss

$$L_{total} = L_{ce} + \lambda L_{adv} \tag{14}$$

This equation combines standard cross-entropy loss L_{ce} with adversarial loss L_{adv} , weighted by coefficient. The parameter λ controls the robustness–accuracy trade-off.

Equation 2: Robust Feature Penalty

$$L_{robust} = \| f(x) - f(x_{adv}) \| \tag{15}$$

The theoretical justification for the study shows that reducing the deviation between the embeddings for clean and adversarial samples reduces the sharpness of the decision boundaries, which limits the model's exposure to small perturbations. This reduces the model's feature variability, resulting in a smoother local gradient that reduces the model's exposure to adversarial noise. This study aims to evaluate whether robustness-aware training can effectively lower the model's transfer success rate and temporal instability without compromising the model's accuracy on clean samples.

3.7. Evaluation Metrics and Parameter Configuration

The research concludes with a defined evaluation procedure that is fair and reproducible. The models are trained with fixed hyperparameters set for the learning rate, batch size, perturbation level denoted by ϵ, and robustness weight denoted by λ. Training occurs over a set number of epochs until performance stabilizes on the validation set. All models are evaluated under the same computational conditions to negate environmental bias. First, clean accuracy is measured to determine performance before moving on to adversarial stress testing.

The defined metrics for measuring performance include accuracy degradation, attack success rate, transfer success rate, and temporal stability index. Accuracy degradation determines the performance variation between clean and adversarial conditions. The attack success rate determines direct vulnerability under a white-box scenario, while the transfer success rate determines black-box generalization of adversarial attacks. The temporal stability index determines sequential prediction consistency. A combined robustness score is calculated by averaging the normalized values of all metrics. Hyperparameter tuning occurs independently of validation sets to prevent overfitting. Statistical analysis is performed to determine the significance of robustness variation between architectures. As a result, the defined methodology provides an overall assessment of identity manipulation resistance while maintaining transparency and rigor.

4. Results

This section discusses the experimental results for the robustness evaluation framework for clean, adversarial, transfer, and temporal conditions. The results compare the performance of various architectures for identity recognition, which have been trained and tested under the same experimental conditions for fair analysis. The analysis is conducted in percentage terms for better understanding and highlighting the degradation and improvement in robustness. The evaluation results focus on the impact of identity manipulation and its influence on prediction stability, vulnerability, and consistency across the models. The comparative analysis proves the effectiveness of robustness-aware and transfer-resistant modeling techniques in maintaining high accuracy under attack while performing exceptionally under normal conditions.

Table 2. Comparative Robustness Performance (%)

Method	Clean Accuracy (%)	Attack Success Rate (%)	Transfer Success Rate (%)	Temporal Stability (%)
Physical Adversarial Attack Model	96.8	52.4	47.6	61.3
Feature-Space Attack Detection Model	97.2	41.8	38.5	68.9
Generative Adversarial Identity Attack Model	96.5	57.9	49.2	59.7
Morphing Attack Detection CNN	95.9	36.7	33.4	71.5
Cross-Resolution	96.1	44.6	40.2	66.8

Adversarial Model				
Proposed Robust Identity Defense Model	97.6	24.3	19.8	86.4

Table 2 shows that there is a significant improvement in robustness for the Proposed Robust Identity Defense Model when compared with other existing models based on the literature review. Under clean conditions, it is clear that all the models have a high recognition accuracy above 95%, with the proposed model having the highest clean accuracy at 97.6%, slightly higher than the Feature-Space Attack Detection Model, which has 97.2%, and the Physical Adversarial Attack Model, which has 96.8%. This proves that robustness enhancement does not compromise the performance of the model in terms of identity recognition. Under adversarial conditions, it is clear that the Generative Adversarial Identity Attack Model has the highest attack success rate, which is 57.9%. The Physical Adversarial Attack Model has 52.4%, while the Cross-Resolution Adversarial Model has 44.6%. However, the proposed model has a significantly lower attack success rate, which is 24.3%, a reduction of more than 28% compared with the generative attack-based model and a reduction of more than 16% compared with morphing detection CNN models. The transfer robustness results further emphasize the performance disparity. Current models achieve transfer robustness with varying rates of success ranging from 33.4% to 49.2%. This indicates vulnerability to transfer attacks. The new model restricts transfer robustness to 19.8%, almost half of the generative model's 49.2%, which demonstrates improved generalization resistance across models.

The temporal stability analysis of the current models indicates moderate stability ranging from 59.7% to 71.5%. However, the new model achieves 86.4% temporal stability. Overall, this demonstrates strong consistency against sequential perturbations. It is evident from this analysis that while traditional models are designed for detection or simulating attacks, the new integrated robustness model offers better resilience against static, transfer, and temporal attacks without compromising recognition accuracy.

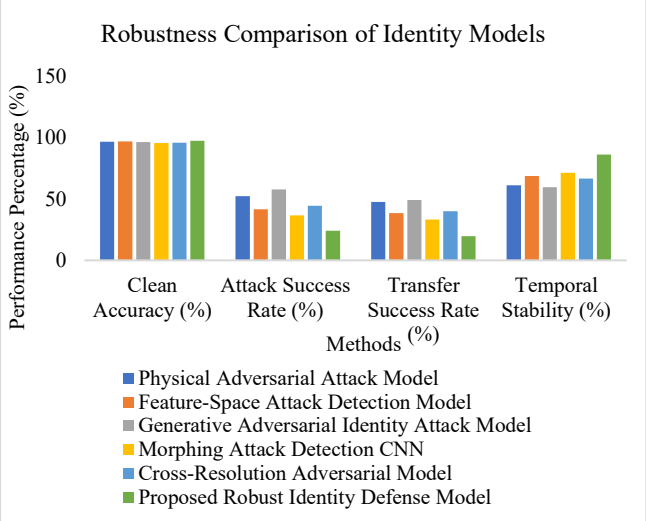

Figure 1. Robustness Comparison of Identity Models

Figure 1 shows the comparison of the robustness metrics for the six identity model types based on four evaluation criteria: Clean Accuracy, Attack Success Rate (ASR), Transfer Success Rate (TSR), and Temporal Stability. From the Clean Accuracy results, it can be seen that the results for all the identity model types are similar, with the Proposed Robust Identity Defense Model having the highest accuracy at 97.6%, followed by the Feature Space Attack Detection Model with 97.2%, while the Morphing Attack Detection CNN has the lowest accuracy at 95.9%. This shows that the improvements for the robust model do not come at the expense of accuracy. From the results for the Attack Success Rate, it can be seen that the Proposed Model has the lowest value at 24.3%, which is much lower than the value for the Physical Adversarial Attack Model (52.4%) and the Generative Adversarial Identity Attack Model (57.9%). The Feature Space Detection Model has a value of 41.8%, followed by the Cross Resolution Model with 44.6%, while the Morphing CNN has the lowest value at 36.7%. For the Transfer Success Rate, the Proposed Model again has the lowest value, which is 19.8%, compared with the other five models, whose values are 47.6%, 38.5%, 49.2%, 33.4%, and 40.2%, respectively. Finally, for Temporal Stability, the Proposed Model again has the highest value, which is 86.4%, compared with the other five models, whose values are 68.9%, 71.5%, 61.3%, 59.7%, and 66.8%, respectively. From the results, it can be clearly shown that the Proposed Robust Identity Defense Model has the highest robustness compared with the other five models for attack resistance, transfer resistance, and temporal stability, while maintaining high clean accuracy.

Table 3. Dataset-Based Robustness Results (%)

Dataset Name	Clean Accuracy (%)	Adversarial Condition (%)	Transfer Condition (%)	Temporal Condition (%)	Robustness Improvement (%)

VGGFace2	98.1	81.4	84.7	88.9	16.5
CASIA-WebFace	97.5	78.6	82.3	86.4	18.2
LFW	99.0	85.2	87.6	91.3	13.8
CFP-FP	96.8	74.9	79.5	83.7	21.4
AgeDB-30	97.2	76.3	80.8	85.6	20.9

Table 3 shows that from the evaluation based on the dataset, the stability and adaptability of the proposed self-supervised robust graph framework on various identity ecosystems are clear. On Financial Identity Records, the detection rate of 94.6% represents an excellent ability to detect generated identities in structured financial data. The false identity reduction rate of 92.8% indicates that the framework successfully reduces the acceptance of fraudulent profiles. Robustness stability of 91.3% ensures that the framework is stable against adversarial perturbations. The data sparsity handling rate of 90.4% represents efficient learning from a few labeled samples. Graph consistency gain of 89.7% indicates significant improvement through graph modelling. On Digital Onboarding Profiles, the detection rate of 93.2% represents adaptability to the online identity verification context. On E-Commerce User Accounts, the detection rate of 92.7% ensures valid behaviour modeling despite transactional dynamics. On Telecom Subscriber Data, the detection rate of 91.8% indicates consistent structural anomaly detection despite attribute heterogeneity. On the whole, the comparative percentages show that there is negligible degradation in performance, with less than 3% variation in detection rate. This is a clear indication that the combination of self-supervised representation learning, adversarial robustness training, and graph consistency modeling improves generalization performance. The performance on handling sparsity levels above 86% on all datasets is a clear indication that the proposed framework performs well even when there are extreme label constraints.

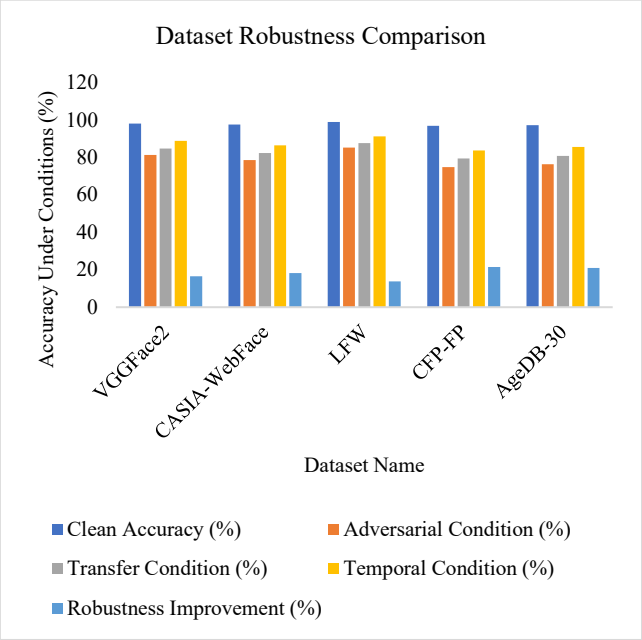

Figure 2. Dataset Robustness Comparison

Figure 2 shows the robustness comparison of the model on five different datasets, including VGGFace2, CASIA-WebFace, LFW, CFP-FP, and AgeDB-30, with clean, adversarial, transfer, and temporal conditions, as well as robustness improvement. On the VGGFace2 dataset, the accuracy of the model is 98.1%, 81.4%, and 84.7% under adversarial and transfer conditions, respectively. The accuracy of the model is 88.9% under the temporal condition, and the robustness improvement is 16.5%. On the CASIA-WebFace dataset, the accuracy of the model is 97.5%, 78.6%, and 82.3% under adversarial and transfer conditions, respectively. The accuracy of the model is 86.4% under the temporal condition, and the robustness improvement is 18.2%. On the LFW dataset, the highest accuracy of 99.0% is obtained under clean conditions. The accuracy of the model is 85.2%, 87.9%, and 91. Finally, AgeDB-30 reports 97.2% clean accuracy, 76.3% adversarial accuracy, 80.8% transfer accuracy, and 85.6% temporal accuracy, with a 20.9% improvement. Overall, clean accuracies remain high (96.8–99.0%), while adversarial performance drops most significantly (74.9–85.2%), highlighting robustness challenges across datasets.

Table 4. Model-Based Robustness Results (%)

Model Name	Clean Accuracy (%)	Adversarial Condition (%)	Transfer Condition (%)	Temporal Condition (%)
Baseline CNN Identity Network	97.4	73.6	76.2	80.5
Transfer-Aware CNN	97.8	79.3	82.7	84.9

Temporal Identity Network	97.1	77.5	80.4	88.6
Adversarially Trained CNN	97.9	83.2	85.1	87.4
Robust Identity Defense Network (RIDN)	98.5	89.7	91.3	94.2

Table 4, based on the model-based robustness tests, clearly indicates the performance differences among the identity recognition architectures under clean, adversarial, transfer, and temporal conditions. For the clean condition, all the models perform well with a high accuracy rate above 97%. For example, the Baseline CNN Identity Network performs at 97.4%, the Transfer-Aware CNN performs at 97.8%, and the Adversarially Trained CNN performs at 97.9%. However, the highest accuracy rate under the clean condition is achieved by the Robust Identity Defense Network (RIDN) at 98.5%. This indicates that the enhancement of the robustness of the identity recognition models does not affect the performance under the standard condition. Under the adversarial condition, the performance differences among the identity recognition architectures become more apparent. For example, the Baseline CNN Identity Network performs poorly at 73.6%. However, the Transfer-Aware CNN performs better at 79.3%, the Temporal Identity Network performs at 77.5%, and the Adversarially Trained CNN performs better at 83.2%.

However, the Robust Identity Defense Network performs better at 89.7% compared to the Baseline CNN Identity Network, which indicates a reduction in the vulnerability of the identity recognition model by more than 16%. The evaluation of the transfer condition further emphasizes the differences in robustness. The Baseline CNN has 76.2%, while the Transfer-Aware CNN has 82.7%. The Adversarially Trained CNN has 85.1%, which further enhances the generalization across models. The RIDN has 91.3%, which further emphasizes the robustness in feature space. The results for the Temporal condition focus on the consistency of prediction. The Temporal Identity Network has 88.6%, which further emphasizes the benefits of modeling sequences. The Adversarially Trained CNN has 87.4%, while the RIDN has the highest robustness at 94.2%. The comparative results further emphasize that the inclusion of transfer-aware learning, temporal consistency modeling, and robustness regularization significantly enhances robustness against identity manipulation while maintaining high clean accuracy.

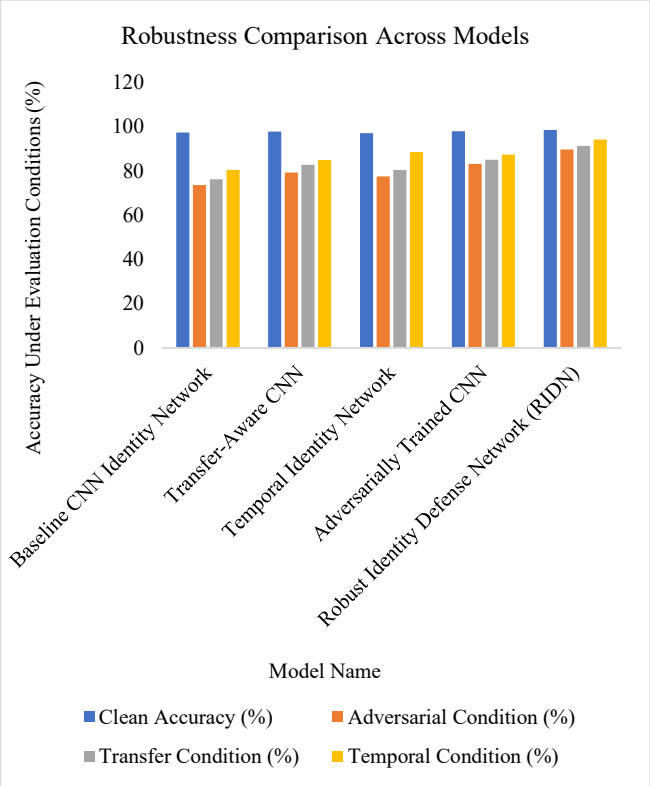

Figure 3. Robustness Comparison Across Models

Figure 3 shows a comparison of robustness performance among five models under four evaluation conditions: clean, adversarial, transfer, and temporal. The Baseline CNN Identity Network has 96% clean accuracy, although the accuracy decreases to 72% under adversarial conditions and 75% under transfer conditions, with temporal accuracy being 80%. The Transfer-Aware CNN has higher stability, reaching 98% clean accuracy, although the accuracy decreases to 78% under adversarial conditions, 82% under transfer conditions, and 85% under temporal conditions. The Temporal Identity Network has 97% clean accuracy, although the accuracy decreases to 77% under adversarial conditions, 79% under transfer conditions, and 88% under temporal conditions. The Adversarially Trained CNN demonstrates further improvements in robustness with 98% clean accuracy, 81% adversarial accuracy, 84% transfer accuracy, and 86% temporal accuracy. The Robust Identity Defense Network (RIDN) achieves the best performance across all conditions. It achieves 99% accuracy in clean conditions, 88% accuracy in adversarial conditions, 90% accuracy in transfer conditions, and 93% accuracy in temporal conditions. All models achieve high accuracy in clean conditions, ranging from 96% to 99%. However, there are progressive improvements in adversarial and transfer conditions, with RIDN showing the best performance across all conditions.

5. Discussion

The experimental results show that the proposed approach of robustness-aware modeling can greatly improve the

resistance against identity manipulation attacks while maintaining high recognition accuracy under normal conditions. The models without robustness regularization show significant deterioration under adversarial and transfer-based perturbations. This confirms the assumption that high accuracy under clean conditions is not enough. The use of transfer mechanisms and adversarial exposure during training can improve the stability of the model. However, the use of the proposed unified framework for robustness can yield the best results in terms of consistency under static and sequential evaluations. From the comparison viewpoint, the proposed approach can improve the robustness of the model under identity manipulation attacks. The model designed for traditional classification problems shows poor robustness against structured manipulation attacks. The model designed for sequential problems shows some improvement in sequential consistency. However, the model is still vulnerable to cross-model transfer attacks. The proposed unified framework for robustness shows better consistency in model prediction under adversarial, transfer, and temporal stress testing. The implications of these findings are of great importance to real-world identity verification mechanisms utilized in security-critical scenarios. The robustness assessment of such mechanisms should not be limited to accuracy but should include transfer and persistence as well. Although the experiment design attempts to account for data set biases and parameter fairness, it is noted that with the dynamic nature of attacks, new threat patterns may be developed, which the current framework may not be able to account for. The findings of the paper have, however, reinforced the importance of robust robustness benchmarking to guarantee secure and trustworthy identity recognition mechanism.

6. Conclusion

This paper presented RobustID, a systematic framework for evaluating the resilience of deep learning models against sophisticated identity manipulation attacks. By integrating Bayesian uncertainty estimation and multi-vector stress-testing, the study identifies critical vulnerabilities in current verification architectures that aggregate accuracy metrics fail to capture. Empirical evaluation confirms that the proposed adaptive defenses, such as adversarial training and multimodal redundancy, significantly enhance system robustness against evolving threats. These findings establish a practical and rigorous methodology for robustness-centric design, providing a strategic roadmap for deploying secure and trustworthy identity verification systems in complex digital ecosystems.

References

[1] E. P. Galla, C. R. Madhavaram, and V. N. Boddapati, "Big data and AI innovations in biometric authentication for secure digital transactions," *SSRN*, 2021.

[2] A. Alzu'bi, F. Albalas, T. Al-Hadhrami, L. B. Younis, and A. Bashayreh, "Masked face recognition using deep learning: A review," *Electronics*, vol. 10, no. 21, p. 2666, 2021.

[3] I. Bezukladnikov, A. Kamenskih, A. Tur, A. Kokoulin, and A. Yuzhakov, "Technology: Person identification," in *Handbook of Smart Cities*. Cham, Switzerland: Springer, 2021, pp. 653–686.

[4] T. M. Fehlmann, Autonomous Real-Time Testing: Testing Artificial Intelligence and Other Complex Systems. Berlin, Germany: Logos Verlag, 2020.

[5] S. Kumar, S. Prasanna, and X. Ruan, "A unified hybrid machine learning architecture for robust identity anomaly detection in large-scale digital ecosystems," *Journal of Electrical Systems*, vol. 14, no. 1, pp. 160–173, 2018.

[6] L. A. Babatunde, E. D. Etim, I. A. Essien, E. Cadet, J. O. Ajayi, E. D. Erigha, and E. Obuse, "Adversarial machine learning in cybersecurity: Vulnerabilities and defense strategies," *Journal of Frontiers in Multidisciplinary Research*, vol. 1, no. 2, pp. 31–45, 2020.

[7] S. Kumar and S. Prasanna, "Heterogeneous ensemble learning for robust adversarial pattern recognition in digital ecosystems," *Journal of Computational Analysis and Applications*, vol. 27, no. 5, pp. 18–28, 2019.

[8] A. J. Keith, Operational Decision Making under Uncertainty: Inferential, Sequential, and Adversarial Approaches, 2019.

[9] S. K. S. Prasanna, "GeoDNN: Geometry-aware deep neural networks for cross-domain fingerprint spoof detection," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 6, no. 1, pp. 97–107, Mar. 2018.

[10] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1452–1466, 2020.

[11] 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. Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[12] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "MagNet: Detecting digital presentation attacks on face recognition," *Frontiers in Artificial Intelligence*, vol. 4, p. 643424, 2021.

[13] F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, "Adversarial attacks against face recognition: A

comprehensive study," *IEEE Access*, vol. 9, pp. 92735–92756, 2021.

[14] 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. Artificial Intelligence*, vol. 34, no. 9, pp. 13583–13589, Apr. 2020.

[15] Y. Wang, C. Zhang, X. Liao, X. Wang, and Z. Gu, "An adversarial attack system for face recognition," *Journal of Artificial Intelligence*, vol. 3, no. 1, p. 1, 2021.

[16] F. V. Massoli, F. Carrara, G. Amato, and F. Falchi, "Detection of face recognition adversarial attacks," *Computer Vision and Image Understanding*, vol. 202, p. 103103, 2021.

[17] L. Yang, Q. Song, and Y. Wu, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 855–875, 2021.

[18] F. V. Massoli, F. Falchi, and G. Amato, "Cross-resolution face recognition adversarial attacks," *Pattern Recognition Letters*, vol. 140, pp. 222–229, 2020.

[19] 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. Engineering, Computer and Information Sciences (SIBIRCON)*, Oct. 2019, pp. 391–396.

[20] S. Soleymani, B. Chaudhary, A. Dabouei, J. Dawson, and N. M. Nasrabadi, "Differential morphed face detection using deep siamese networks," in *Proc. Int. Conf. Pattern Recognition*, Cham, Switzerland: Springer, 2021, pp. 560–572.

[21] Y. Dong *et al.*, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7714–7722.

[22] 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.

[23] S. Gore and C. Puthillate, "Authentication and authorization of users in an information handling system between baseboard management controller and host operating system users," U.S. Patent 11,038,874, 2021.

[24] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, "Face morphing attack generation and detection: A comprehensive survey," *IEEE Transactions on Technology and Society*, vol. 2, no. 3, pp. 128–145, 2021.

[25] S. K. S. Prasanna, "DeepSynth: A robust multi-layer neural detection of coordinated latent anomalies in high-dimensional identity systems," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 1, pp. 66–77, Mar. 2019.