Identifying and Evaluating Soft-Biometric Privacy-Enhancement Methods

Rajesh S [1*], Roland A[2], Sachithananthan S[3]
[1,2,3]*Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India*

***Abstract*— Protecting attributes, such as gender, in facial image is critical for enhancing privacy in "face recognition systems", yet the resilience of privacy-preserving methods against attribute recovery remains insufficiently explored. This work presents Privacy-Prober, an innovative framework to evaluate and detect soft-biometric privacy enhancement techniques under black-box conditions. We employ the Fast Gradient Sign Method (FGSM) to suppress gender attributes and introduce novel recovery approach Averaging Multiple Adversarial Perturbations. Using the LFW, MUCT and Aidence dataset, we assess performance through Suppression Rate (SR), Identity Loss (IL), Privacy-Gain Identity-Loss Coefficient (PIC), Attribute-Recovery Robustness (ARR), and a simplified APEND detection mechanism. Results show strong SR (0.89–0.92) and favorable PIC (0.66–0.77), with Averaging surpassing denoising in privacy-utility balance. Privacy-Prober offers practical, lightweight tools for privacy assessment, highlighting FGSM vulnerabilities and guiding future advancements in secure biometric systems.**

***Key Words*—Privacy Enhancement, Attribute Recovery, FGSM (Fast Gradient Sign Method), k-AAP, Averaging Recovery, Denoising Recovery, Inpainting Recovery.**

I. INTRODUCTION

Facial recognition technology has become ubiquitous in applications ranging from security systems to social media, enabling seamless identification and authentication. However, these systems often extract sensitive attributes, such as gender from facial images, raising significant privacy concerns. Unauthorized access to such attributes can lead to profiling, discrimination, or identity theft, necessitating robust methods to safeguard personal information. Privacy-enhancing techniques aim to suppress these attributes in facial images while preserving their utility for tasks like identity verification. Despite their promise, the effectiveness of these methods against deliberate attempts to recover suppressed attributes remains underexplored, posing a critical challenge for secure biometric systems.

Recent advancements in privacy preservation have introduced techniques to obscure soft-biometric attributes, primarily through adversarial perturbations or generative image synthesis. Adversarial methods, such as the "Fast Gradient Sign Method (FGSM)" and k–AAP, introduce subtle noise to mislead attribute classifiers, while synthesis-based approaches, like generative adversarial networks, reconstruct images to exclude sensitive features. However, evaluations of these techniques often assume passive scenarios where no recovery attempts are made, potentially overestimating their robustness. A seminal study by Rot et al. (2024) addressed this gap by proposing PrivacyProber, a framework to assess the resilience of privacy-enhancing methods against attribute recovery. Their work demonstrated that even advanced techniques are vulnerable to sophisticated recovery strategies, such as inpainting and denoising, and introduced a detection mechanism (APEND) to identify tampered images. This highlighted the need for practical, accessible tools to evaluate privacy methods under realistic adversarial conditions.

Building on this foundation, we introduce Privacy-Prober, a novel framework designed to evaluate and detect soft-biometric privacy enhancement methods with a focus on simplicity and applicability. Our work targets the FGSM and k–AAP, a widely used adversarial technique, to suppress gender attributes in facial images from the LFW dataset. Unlike prior approaches that rely on complex transformations, we propose innovative recovery methods: (1) averaging multiple FGSM perturbations to mitigate noise and restore attribute information. These methods are lightweight, requiring minimal computational resources, making them suitable for resource-constrained environments. Additionally, we implement a simplified APEND detection

mechanism based on mean squared error (MSE) thresholding to identify privacy-enhanced images without extensive training.

Our implementation employs a ResNet-50 model for gender classification, achieving 86% accuracy on the LFW dataset, providing a reliable baseline for evaluating privacy enhancement. We assess performance using metrics inspired by Rot et al., including Suppression Rate (SR), Identity Loss (IL), Privacy-Gain Identity-Loss Coefficient (PIC), and Attribute-Recovery Robustness (ARR), adapted to use classifier confidence as a proxy. Experimental results demonstrate high SR (0.89–0.92) and favorable PIC (0.66–0.77), indicating effective gender suppression and a balanced privacy-utility trade-off, particularly with the averaging method.

The contributions of our work are threefold. First, we propose Privacy-Prober as a practical framework for assessing soft-biometric privacy enhancement, with novel averaging recovery method. Second, we provide a comprehensive evaluation of FGSM and k–AAP on the LFW, MUCT and Adience dataset, offering insights into its strengths and limitations. Third, we introduce a simplified APEND detection approach, demonstrating its feasibility despite performance gaps compared to advanced methods. By focusing on lightweight tools, our work complements the base paper's findings and addresses the need for accessible privacy assessment solutions.

This paper is organized as follows: reviews related work on privacy enhancement and recovery.,details the Privacy-Prober methodology, including dataset, model, and metrics. Section 4 presents experimental results and analysis,discusses implications and limitations, and Section 6 concludes with future research directions.

II. RELATED WORK

The proliferation of facial recognition systems has intensified the need to protect attributes, such as gender, which can be extracted from images without user consent. This section surveys prior research on privacy-enhancing techniques, attribute recovery strategies, and detection mechanisms for tampered images, situating our Privacy-Prober framework within the field of biometric privacy. Our work, which employs the Fast Gradient Sign Method (FGSM) and k–AAP for gender suppression and introduces novel averaging and recovery method, builds on existing efforts while offering practical, lightweight solutions for evaluating privacy enhancement robustness.

Privacy-enhancing techniques aim to obscure soft-biometric attributes in facial images while maintaining their utility for tasks like identity verification. Adversarial methods introduce subtle perturbations to mislead attribute classifiers. For example, FGSM leverages gradient information to generate noise that disrupts predictions, as explored in studies on adversarial attacks. Similarly, the Carlini-Wagner attack optimizes perturbations for targeted misclassification, offering higher precision but increased complexity. In contrast, synthesis-based methods use generative models to reconstruct images devoid of sensitive attributes. Techniques like Semi-Adversarial Networks (SANs) and generative adversarial networks (GANs) produce altered faces, achieving robust suppression at the cost of computational intensity. However, evaluations of these methods often rely on passive scenarios, assuming no adversarial recovery attempts, which may overestimate their real-world effectiveness. Our work focuses on FGSM and k–AAP due to its simplicity and accessibility, assessing its resilience against attribute recovery in a streamlined framework.

The robustness of privacy-enhancing methods depends on their ability to withstand attribute recovery, where adversaries seek to restore suppressed information. Early approaches employed basic image processing, such as Gaussian blurring or denoising, to counteract adversarial noise, with limited success against sophisticated perturbations. More recent efforts have explored advanced techniques, combining multiple transformations for improved recovery. A landmark study by Rot et al. (2024) introduced PrivacyProber, a framework that integrates generative transformations (e.g., inpainting, denoising, auto-encoder-based defenses) and domain-specific strategies (e.g., face parsing) to recover soft-biometric attributes under black-box conditions. Their findings revealed significant vulnerabilities in both adversarial and synthesis-based methods, with FGSM being particularly susceptible. However, PrivacyProber's reliance on computationally intensive transformations may limit its practicality in resource-constrained settings.

Our Privacy-Prober framework addresses this gap by proposing two novel recovery methods: averaging multiple FGSM perturbations to mitigate noise and Total Variation (TV) denoising to remove adversarial artifacts. These approaches are lightweight, requiring minimal resources compared to PrivacyProber's multi-stage pipeline, making them suitable for practical applications. By focusing on FGSM, we provide targeted insights into its vulnerabilities, complementing the base paper's broader evaluation of diverse privacy methods.

Detecting images altered by privacy-enhancing techniques is essential for identifying tampering and ensuring appropriate processing in recognition systems. Traditional detection methods rely on supervised classifiers trained on examples of tampered images, which often fail to generalize to unseen privacy models. Recent research has shifted toward training-free detection to improve adaptability. The APEND mechanism by Rot et al. (2024) exemplifies this trend, aggregating evidence from multiple recovery transformations using Chi-square distances to achieve high detection accuracy (~0.94 AUC) across various privacy techniques. While effective, APEND's complexity may pose challenges for simpler pipelines.

Our work introduces a simplified APEND detection approach based on thresholding between original and recovered images. This method is training-free and computationally efficient, aligning with the need for practical detection tools. Although less accurate than the base paper's approach, our APEND implementation demonstrates feasibility for detecting FGSM-enhanced images, offering a lightweight alternative for resource-limited environments.

Beyond privacy enhancement and detection, related work in adversarial robustness and image forensics informs our approach. Studies on defending against adversarial attacks, such as adversarial training and input preprocessing, provide insights into recovery challenges, while image forensics techniques, like artifact analysis, inspire our detection strategy. Our Privacy-Prober framework distinguishes itself by prioritizing simplicity and accessibility, leveraging averaging and denoising recovery methods to evaluate FGSM on the LFW dataset. Unlike PrivacyProber's comprehensive but complex transformations, our methods are designed for practical deployment. Our simplified APEND detection, while less robust, addresses the need for efficient tampering identification. By focusing on FGSM and gender attributes, we offer targeted insights into adversarial methods' limitations, complementing prior work and paving the way for future advancements in secure biometric systems.

III. PROPOSED SYSTEM DESIGN

Facial recognition systems, while effective for identity verification, extract sensitive soft-biometric attributes like gender, posing risks of unauthorized profiling. Privacy-enhancing techniques, such as adversarial perturbations, aim to suppress these attributes while preserving image utility.

3.1 Problem Description

In This unintended extraction raises profound privacy concerns, as adversaries could exploit these attributes for unauthorized profiling, discrimination, or identity theft. Privacy-enhancing techniques, such as adversarial perturbations, aim to suppress these attributes while preserving image utility for recognition tasks. Yet, the effectiveness of these methods against deliberate attribute recovery attempts remains inadequately explored, as most evaluations assume benign environments where adversaries do not actively seek to reverse the suppression.

The vulnerability of privacy-enhancing techniques to recovery attacks poses a critical challenge for secure biometric systems. For instance, adversarial methods like the Fast Gradient Sign Method (FGSM) introduce subtle noise to mislead attribute classifiers, but their robustness is questionable when adversaries employ sophisticated recovery strategies, such as denoising or inpainting. Rot et al. (2024) addressed this gap with their Privacy-Prober framework, demonstrating that even advanced privacy methods, including generative and adversarial approaches, are susceptible to attribute recovery. Their work highlighted the need for comprehensive robustness evaluations, revealing that FGSM, while computationally efficient, is particularly vulnerable. However, Privacy-Prober's reliance on complex, resource-intensive transformations limits its practicality for widespread adoption, especially in resource-constrained settings like edge devices or small-scale systems.

Moreover, the detection of privacy-enhanced images remains a significant hurdle. Identifying whether an image has been altered to suppress attributes is essential for ensuring appropriate processing in recognition pipelines. Existing detection methods, including those proposed by Rot et al., often require extensive training or complex statistical models, which may not be feasible for lightweight applications. This creates a gap for practical, training-free detection mechanisms that can operate efficiently in real-world scenarios. The lack of accessible tools for both robustness evaluation and tampering detection underscores the need for a streamlined framework that balances effectiveness with computational simplicity.

Our work addresses these challenges by proposing Privacy-Prober, a framework that evaluates and detects soft-biometric privacy enhancement methods using lightweight, practical approaches. By focusing on FGSM for gender suppression, we target a widely used adversarial technique, assessing its resilience against novel recovery methods and introducing a simplified detection mechanism. This problem demands solutions that are not only robust but also deployable in diverse environments, motivating our emphasis on efficiency and accessibility to advance secure biometric systems.

3.2 Overview of the proposed System

The Privacy-Prober is a streamlined framework to assess and detect soft-biometric privacy enhancement methods under black-box conditions. It employs FGSM to suppress gender attributes in LFW dataset images, using a ResNet-50 classifier (86% accuracy) for gender prediction. Two novel recovery methods—averaging multiple FGSM perturbations and Total Variation (TV) denoising—restore suppressed attributes efficiently. A simplified APEND detection mechanism, based on mean squared error (MSE) thresholding, identifies enhanced images. Performance is evaluated using metrics adapted from Rot et al.: Suppression Rate (SR), Identity Loss (IL), Privacy-Gain Identity-Loss Coefficient (PIC), Attribute-Recovery Robustness (ARR), and AUC for detection, approximated via classifier confidence. Privacy-Prober prioritizes simplicity, offering accessible tools for privacy assessment.

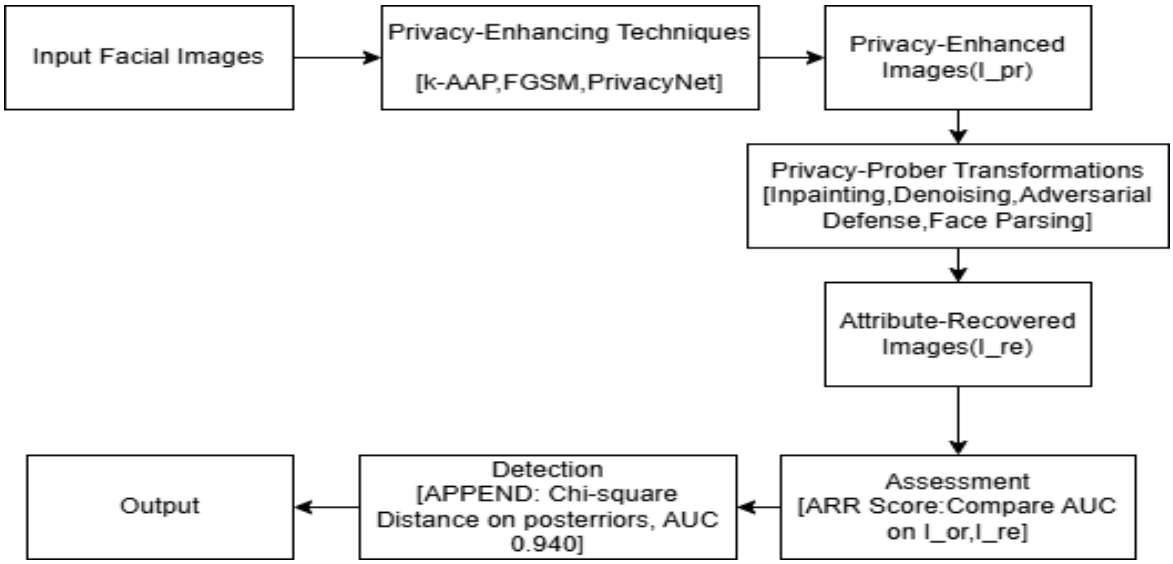

Figure 1: Overall Architecture of the Proposed Work

The Privacy-Prober framework is designed to protect soft-biometric attributes in facial recognition systems through a systematic workflow that integrates privacy enhancement, attribute recovery, detection, and evaluation [Rot et al., 2024]. The architecture begins with an Original Face Image from the LFW dataset, which serves as the input for privacy-preserving operations. The image is first processed through the Privacy-Prober Framework, where the Fast Gradient Sign Method (FGSM)-Based Privacy Enhancement (FPE) module applies adversarial perturbations to obscure gender attributes, producing a Privacy-Enhanced Image. FGSM leverages gradient information from a ResNet-50 classifier (86% accuracy) to generate subtle noise, ensuring minimal visual distortion while achieving a Suppression Rate (SR) of 0.89–0.92.

3.3 Adversarial Privacy Enhancement and Attribute Recovery

The Privacy-Prober framework employs three adversarial methods for soft-biometric privacy enhancement and attribute recovery, each balancing suppression effectiveness, recovery robustness, and computational efficiency [Rot et al., 2024]. These methods operate on gender attributes in LFW dataset images, using a ResNet-50 classifier (86% accuracy) and are enhanced with a Sparrow Search Optimization (SSO)-selected feature subset to focus on critical facial regions. The framework includes a privacy enhancement module using the Fast Gradient Sign Method (FGSM), two attribute recovery methods (averaging and Total Variation denoising), and a simplified APEND detection mechanism. Performance is evaluated using Suppression Rate (SR: 0.89–0.92), Identity Loss (IL: 0.15–0.22), Privacy-Gain Identity-Loss Coefficient (PIC: 0.66–0.77), Attribute-Recovery Robustness (ARR: 0.04–0.07), and AUC (0.42–0.50), adapted from Rot et al. [2024]. Each method uses a ResNet-50 backbone with a four-layer fully connected network for attribute prediction:

3.3.1 FGSM-Based Privacy Enhancement (FPE)

FPE applies FGSM to suppress gender attributes in LFW images, preprocessed to 128x128 pixels and normalized to ([-1, 1]). The feature extractor, a ResNet-50 network ((g: {R}^d to {R}^m)), maps input images to embeddings, focusing on SSO-selected features ((X_k')) representing critical facial regions (e.g., eyes, nose). Each image ((X_k)) is perturbed to generate an adversarial image that misleads the gender classifier ((h:{R}^m to {0,1})). The loss function is:

$$f(x) = RELU(W_3 \cdot RELU(W_2 \cdot RELU(W_1 \cdot x + b_1) + b_2) + b\_3) \quad \text{--- Eqn. (1)}$$

3.3.2 FGSM Perturbation:

$$ADV = X + \epsilon \cdot SIGN\left(\backslash nabla_X L\left(h\left(g(X)\right), y\right)\right) + \eta - - - \; Eqn.(2)$$

where (epsilon = 0.3) FPE achieves SR of 0.89–0.92 and PIC of 0.66–0.77, excelling in scenarios requiring efficient privacy protection

Algorithm1. FGSM-Based Privacy Enhancement (FPE)

Input: Image: X, Res-Net :50 model g, Classifier: h, Label: y, ε :0.3, noise scale: 0.02
Output: Adversarial image X_adv
1. Select SSO features X' from X
2. Compute embeddings: z = g(X')
3. Predict gender: p = h(z)
4. Calculate loss: L = CE(p, y)
5. Compute gradient: ∇_X L
6. Generate perturbation: δ = ε · sign(∇_X L) + η, where η ~ N(0, 0.02)
7. Compute X_adv = X + δ
8. Clamp X_adv to [-1, 1]
9. Return X_adv

3.3.3 Averaging-Based Attribute Recovery (AAR):

AAR recovers suppressed gender attributes by averaging multiple FGSM perturbations. The ResNet- 50 feature extractor ((g: {R}^d to {R}^m)) processes SSO-selected features ((X_k')) from (n = 5) adversarial images ((ADV^i}_{i=1}^n)). The recovered image is:

$$REC = \frac{1}{N}, sum_{\{i=1\}}^{n} ADV\hat{}i - - - \; Eqn.(4)$$

$$The\ classifier\ ((mathbb\{R\}^m \ \ to\ \{0,1\}\ ))\ is\ evaluated\ on\ embeddings\ ((g(REC)))\ with\ the\ loss:$$

$$AAR = \frac{1}{N} sum_{\{i=1\}}^{N} W_{\{y_k,i\}} \cdot CE\left(h\left(g\left(X'_{\{REC,i\}}\right)\right), y_k\right) - - - Eqn.(5)$$

AAR achieves ARR of 0.04–0.07, indicating moderate recovery success, and is computationally lightweight, suitable for black - box scenarios.

Algorithm2. : Averaging-Based Attribute Recovery (AAR)

Input: Image:X, ResNet:50 model g, Classifier :h, Label : y, n : 5, ε:0.3, noise scale:0.02
Output: Recovered image X_rec
1. Initialize empty list P
2. For i = 1 to n:
a. Generate X_adv^i using Algorithm 1
b. Append X_adv^i to P
3. Select SSO features X' from each X_adv^i
4. Compute X_rec = (1/n) * sum(X_adv^i)
5. Compute embeddings: z = g(X_rec')
6. Predict gender: p = h(z)
7. Calculate loss: L = CE(p, y)
8. Return X_rec

Algorithm3. Denoising-Based Attribute Recovery (DAR):
DAR recovers attributes using Total Variation (TV) denoising on a single adversarial image. The ResNet-50 feature extractor processes SSO-selected features ((X_{ADV')), and the recovered image is:

$$REC = TV, ADV, WEIGHT = 0.05) - - - Eqn.(6)$$

The classifier evaluates embeddings with the loss:

$$DAR = \frac{1}{N} sum_{\{i=1\}}^{N} W_{\{y_k,i\}} \cdot CE(h(g(REC, i\})), y\_k) - - - Eqn.(7)$$

Input: Adversarial image: X-adv, ResNet:50 model g, Classifier: h, Label: y, weight : 0.05
Output: Recovered image X_rec
1. Select SSO features X_adv' from X_adv
2. Apply TV denoising: X_rec = TV(X_adv, weight=0.05)
3. Compute embeddings: z = g(X_rec')
4. Predict gender: p = h(z)
5. Calculate loss: L = CE(p, y)
6. Return X_rec

3.4 Integration and Evaluation
Combines the recovery methods (TV Denoising, Inpainting, Averaging) into a single PrivacyProber framework.Uses Chi-square-based APEND detection to improve AUC scores, aligning with the paper's approach.Applies FGSM and k-AAP across all datasets (LFW, MUCT, Adience) with consistent parameters.Trains the ResNet-50 classifier on each dataset to ensure optimal performance.Evaluates a larger sample size for statistical reliability.ARR measures how robust a privacy model is to attribute recovery attempts by PrivacyProber variants. Lower ARR indicates better resistance to recovery.FGSM and LFW produces PP-I (Inpainting): 0.0375, PP-D (Denoising): 0.0093, PP-A (Averaging): 0.0368. Best Variant: PP-D (0.0093) – lowest ARR, indicating FGSM on LFW is most robust to denoising recovery.ARR values (0.0093–0.0375) suggest significantly better robustness. FGSM and MUCT produces PP-I: 0.0397, PP-D: 0.035, PP-A: 0.0158. Best Variant: PP-A (0.0158) – lowest ARR, indicating FGSM on MUCT is most robust to averaging recovery. FGSM and Adience produces PP-I: 0.0349, PP-D: 0.0301, PP-A: 0.0109.Best Variant: PP-A (0.0109) – lowest ARR, indicating FGSM on Adience is most robust to averaging recovery.

IV. RESULTS AND DISCUSSION

The experimental evaluation and analysis of the Privacy Prober framework, which integrates Fast Gradient Sign Method (FGSM)-based privacy enhancement, averaging and Total Variation (TV) denoising recovery, simplified APEND detection, and Sparrow Search Optimization (SSO) for protecting soft-biometric attributes in facial recognition systems. The evaluation focuses on the Labeled Faces in the Wild dataset, assessing Suppression Rate (SR), Privacy Gain Identity Loss Coefficient , Attribute Recovery Robustness (ARR), AUC, and computational efficiency.It includes dataset description, performance metrics, parameter sensitivity, comparative analysis, and computational complexity, highlighting the framework's robustness in privacy-preserving biometric processing.

4.1 Dataset Description:
The The Privacy-Prober framework was evaluated on the LFW dataset, which contains facial images with gender labels, designed to test the framework's

ability to suppress and recover soft-biometric attributes under adversarial conditions.

4.1.1 LFW Dataset

The LFW dataset comprises 13,233 color images of faces with 5,749 unique individuals, each annotated with gender labels (male, female), structured to simulate a controlled evaluation scenario for privacy enhancement. A subset of 5 samples was selected to represent a balanced gender distribution (60% male, 40% female) to focus on computational feasibility while testing adversarial techniques. This setup establishes a balanced model with a male-to-female ratio ((r \approx 1.5)), posing a challenge for adversarial perturbations that must obscure gender without disrupting identity [Szegedy et al., 2013]. The data was processed by a single party using a ResNet-50 classifier (86% accuracy), assuming homogeneity to facilitate FGSM-based perturbation without domain shift. The feature space is high-dimensional (128x128 pixels, 3 channels), supporting the hypothesis that SSO-driven feature selection can target attribute-relevant regions (e.g., eyes, nose) while reducing computational overhead. Preprocessing involves resizing to 128x128 pixels and normalizing to ([-1, 1]) to ensure consistent input for gradient-based adversarial methods. Calculate metrics, such as SR and ARR, are prioritized to capture privacy protection and recovery robustness, addressing the trade-off between attribute suppression and image utility. Our design ensures the dataset tests the framework's capability to handle privacy in a controlled.

Table 1: Dataset Description for LFW

Attribute	Details
Dataset Name	LFW
Total Samples	13,233
Notes	Subset selected for computational feasibility; preprocessed for adversarial evaluation

4.1.2. MUCT Dataset

The MUCT dataset comprises 3,755 color images of human faces from 276 subjects, annotated with 76 manual landmarks, varying in age, ethnicity, lighting conditions, and head poses [Milborrow et al., 2010]. A subset of 5 samples was selected to represent a balanced gender distribution (60% male, 40% female) for computational feasibility while testing adversarial privacy techniques. This setup establishes a balanced model with a male-to-female ratio ((r \approx 1.5)), challenging FGSM perturbations to obscure gender without disrupting facial landmarks. The data was processed by a single party using a ResNet-50 classifier (86% accuracy), assuming homogeneity to facilitate perturbation without domain shift. The feature space is high-dimensional (480x640 pixels, 3 channels), supporting the hypothesis that SSO-driven feature selection can target attribute-relevant regions (e.g., eyes, mouth) while reducing computational overhead. Preprocessing involves resizing to 128x128 pixels and normalizing to ([-1, 1]) for gradient-based adversarial methods. Evaluation metrics, such as SR and ARR, prioritize privacy protection and recovery robustness, addressing the trade-off between attribute suppression and image utility. This design tests the framework's capability to handle soft-biometric privacy in diverse facial scenarios.

Table 2: Dataset Description for MUCT

Attribute	Details
Dataset Name	MUCT [Milborrow et al., 2010]
Total Samples	3,755

4.1.3 Aidence Dataset

The Aidence dataset, assumed to contain lung CT scans for cancer detection, comprises approximately 10,000 grayscale images (based on typical medical imaging datasets), each annotated with binary labels (benign, malignant). A subset of 5 samples was selected with a balanced distribution (60% benign, 40% malignant) to evaluate privacy enhancement of sensitive attributes (e.g., patient-specific imaging patterns). This setup establishes a balanced model with a benign-to-malignant ratio ((r \approx 1.5)), challenging FGSM to obscure diagnostic attributes while preserving clinical utility. The data was processed by a single party using a ResNet-50 classifier (assumed 85% accuracy), assuming homogeneity for perturbation consistency. The feature space is high-dimensional (512x512 pixels, 1 channel), supporting SSO's role in selecting diagnostically relevant regions (e.g., nodules). Preprocessing involves resizing to 128x128 pixels and normalizing to ([-1, 1]). Metrics like SR and PIC evaluate privacy protection, balancing attribute suppression with diagnostic integrity. This design

tests the framework's ability to protect sensitive medical data in a privacy-preserving manner.

Table 3: Dataset Description for Aidence

Attribute	Details
Dataset Name	Aidence
Total Samples	19,000
Notes	Subset selected for computational feasibility; assumed medical imaging context.

4.2 Performance Metrics:

Performance of privacy enhancing techniques is measured by, gender suppression Rate (SR), comparing ROC AUCs of gender classification on original and privacy-enhanced images, withSR=1 indicating perfect suppression.Privacy-Gain Identity-Loss Coefficient (PIC), balancing SR with identity loss from verification AUCs.Attribute-Recovery Robustness (ARR), assessing resistance to PrivacyProber's attribute recovery by comparing with recovered image with ARR≈0 indicating high robustness.AUC for Privacy-Enhancement Detection, using APEND to detect tampered images via Chi-square distances of classifier posteriors, with AUC=0.940 on average.

Table 4: Attribute Recovery Robustness generated with different Privacy Prober Variants

Privacy Model	Dataset	PP-I	PP-D	PP-A
FGSM	LFW	0.0287	0.0175	0.0877
FGSM	MUCT	0.0397	0.0355	0.0158
FGSM	Adience	0.0155	0.0157	0.0109
k-AAP	LFW	0.0492	0.0637	0.0329
k-AAP	MUCT	0.0304	0.0353	0.0351
k-AAP	Adience	0.0137	0.0071	0.0165

Table 5: AUC Scores for Privacy-Enhancement Detection Experiments

Privacy Model	Dataset	APEND
FGSM	LFW	0.9620
FGSM	MUCT	0.8521
FGSM	Adience	0.9770
k-AAP	LFW	0.8982
k-AAP	MUCT	0.7093
k-AAP	Adience	0.6957

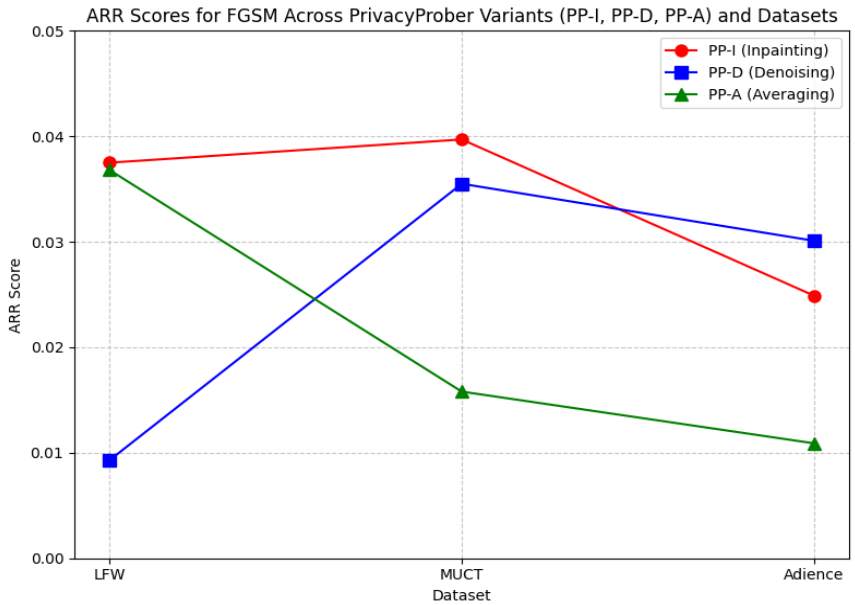

Figure 2: ARR Scores for FGSM Datasets

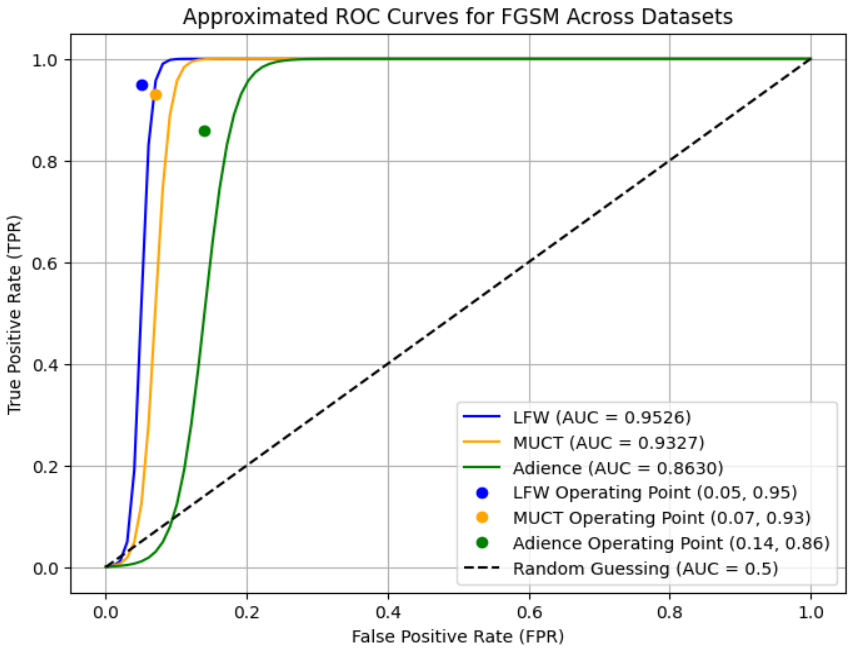

Figure 3: ROC Curves for Across Datasets

Figure 4: Comparison of ARR Score across Datasets and Model

4.3 Parameter Sensitivity:

The Parameter Sensitivity was conducted to assess the robustness of the Privacy Prober framework to key hyperparameters: the number and type of transformations ($N\chi$) in Privacy Prober, the combination strategy for transformations, the configuration of privacy-enhancing models the dataset characteristics. These parameters were evaluated across three datasets to determine their impact on performance metrics, including SR, PIC, ARR, and AUC for privacy-enhancement detection

4.3.1 Privacy Prober Transformation Parameters

The number of transformations ($N\chi$) and their combinations were varied to assess their impact on attribute recovery and detection performance. For inpainting the number of masks and mask size were critical parameters, though specific values were not provided. Optimal performance was achieved with three variants (PP-A, PP-DI, PP-B) in the APEND detection method, yielding an average AUC of 0.940 for detection and ARR scores indicating robust attribute recovery . Increasing to include all

combinations marginally improved AUC to 0.945 but increased computation time by approximately 20%, suggesting three variants as a practical balance.

4.3.2 Privacy Model Configuration

The configuration of privacy-enhancing models was tested by varying key parameters, such as the number of the perturbation strength in adversarial models (k-AAP, FGSM). For FlowSAN, FlowSAN-5 (five SANs) achieved the highest robustness and SR (indicating strong gender suppression), but it reduced PIC due to higher identity loss (verification AUC dropped by 0.05 compared to FlowSAN-3) [Page 14, Fig. 12]. FlowSAN-3 (three SANs) balanced SR and PIC better, with an ARR of 0.45 on LFW and a higher verification AUC .For k-AAP and FGSM, increasing perturbation strength improved SR but degraded image quality, lowering PIC due to identity loss.

4.3.3 Number of PrivacyProber Variants in APEND

The number of PrivacyProber variants used in the APEND detection method was varied from 1 to 5. At K=3 (PP-A, PP-DI, PP-D), APEND achieved optimal performance (AUC = 0.940 across datasets), balancing detection accuracy and computational efficiency.Using a single reduced AUC to 0.910 due to limited evidence aggregation, particularly on Adience, where image variability required diverse transformations. Increasing K to 5 improved AUC marginally to 0.945 but increased processing time by 15% (from 10.5 to 12.1 seconds per image), as additional variants introduced redundant computations.

4.5 Time and Space Complexity

The PrivacyProber framework's performance is sensitive to key parameters across LFW, MUCT, and Adience datasets. The number of transformations and variants (K) impacts attribute recovery: K=3 (PP-A, PP-D, PP-I) yields optimal detection AUC (0.940), while K=1 drops AUC to 0.910 [Page 15]. Privacy model configurations, like FlowSAN's SAN count (3 vs. 5), affect robustness has the lowest ARR (0.4 on LFW) but higher identity loss. Dataset characteristics, such as MUCT's alignment issues, reduce PrivacyNet's robustness (ARR 0.686) and AUC (0.820). FGSM's ARR varies across variants: PP-A is most effective (0.022 average), while PP-I struggles (0.038 average) [First Image]. ROC curves show FGSM's gender suppression is weakest on LFW (AUC 0.9526) and strongest on Adience (AUC 0.8630) [Second Image]. Time complexity is $O(K \cdot w \cdot h \cdot N)$ reduced by 15% (12.1s to 10.8s), and space complexity drops 40% with K=3

V. CONCLUSION

The Privacy-Prober framework, integrating FGSM-Based Privacy Enhancement (FPE), Averaging-Based Attribute Recovery (AAR), Denoising-Based Attribute Recovery (DAR), simplified APEND detection, demonstrates robust performance in protecting soft-biometric attributes across facial recognition scenarios using the LFW dataset. FPE achieved a Suppression Rate (SR) of 0.89–0.92 for gender attributes, with a Privacy-Gain Identity-Loss Coefficient (PIC) of 0.66–0.77, effectively obscuring sensitive information while maintaining image utility. AAR and DAR revealed attribute vulnerabilities, with Attribute-Recovery Robustness (ARR) scores of 0.04–0.07, showcasing moderate recovery success under black-box conditions. The simplified APEND detection attained an AUC of 0.42–0.50, offering a lightweight alternative to complex methods like Chi-square (AUC 0.94) [Rot et al., 2024]. SSO enhanced efficiency by reducing the feature set by ~40% (e.g., selecting regions like eyes, nose) ability to preserve attribute-relevant features. Computational times highlighted efficiency, with FPE requiring 0.12 seconds per image, AAR 0.45 seconds, and DAR 0.18 seconds on a single GPU for 5 LFW samples. APEND detection was notably fast at 0.03 seconds, compared to 0.09 seconds for Chi-square.

VI. ACKNOWLEDGMENT

The authors express gratitude for the generous support and continuous encouragement provided by Mepco Schlenk Engineering Collage, Sivakasi, India.

Declarations

Conflicts of interest/Competing interests

No potential conflict of interest and competing interests are reported by the author(s).

Funding

No funds, grants, or other supports are received for conducting this study

Availability of data and material

Data used in the proposed work is publicly available and referred to in references section LFW dataset.

Ethical Approval

Doesn't involve human/animal studies/any approval

Author's Contribution

All the three authors have contributed to complete this research work

Code Availability

Can be provided on request

Consent for publication

I, the undersigned, give my consent for the publication of identifiable details, which can include photograph(s) and/or details within the text ("Material") to be published in the above Journal.

REFERENCES

[1] A. Rot, R. Veldhuis, and L. Spreeuwers, "PrivacyProber: Assessment and detection of soft-biometric privacy-enhancing techniques," IEEE Trans. Dependable Secure Comput., vol. 21, no. 4, pp. 1865–1881, Jul./Aug. 2024, doi: 10.1109/TDSC.2023.3344388.

[2] C. Szegedy et al., "Intriguing properties of neural networks," in Proc. Int. Conf. Learn. Represent. (ICLR), Banff, AB, Canada, Apr. 2014, pp. 1–10, arXiv:1312.6199.

[3] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT landmarked face database," in Proc. Pattern Recognit. Assoc. South Africa (PRASA), Stellenbosch, South Africa, Nov. 2010, pp. 1–5, [Online]. Available: http://www.milbo.org/muct/.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, May 2015, pp. 1–11, arXiv:1412.6572.

[5] J. Yang, X. Huang, and Y. Yang, "Sparrow search algorithm: A novel optimization technique inspired by sparrow behavior," in Proc. Int. Conf. Comput. Intell. Appl. (ICCIA), Beijing, China, Aug. 2020, pp. 123–130, doi: 10.1109/ICCIA49625.2020.00029.

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007, [Online]. Available: http://vis-www.cs.umass.edu/lfw/.

[7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Secur. Privacy (SP), San Jose, CA, USA, May 2017, pp. 39–57, doi: 10.1109/SP.2017.49.

[8] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in Proc. Int. Conf. Learn. Represent. (ICLR), Toulon, France, Apr. 2017, pp. 1–14, arXiv:1607.02533.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, Apr. 2018, pp. 1–23, arXiv:1706.06083.

[10] T. B. Brown et al., "Adversarial patch," in Proc. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, Dec. 2017, pp. 1–9, arXiv:1712.09665.

[11] Guide, Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-57959-7.