Shielding Latent Face Representations From Privacy Attacks

Arjun Ramesh Kaushik[1], Bharat Chandra Yalavarthi[1],
Arun Ross[2], Vishnu Boddeti[2], Nalini Ratha[1]
[1] University at Buffalo, Buffalo, USA
[2] Michigan State University, East Lansing, USA

Abstract—In today's data-driven analytics landscape, deep learning has become a powerful tool, with latent representations, known as embeddings, playing a central role in several applications. In the face analytics domain, such embeddings are commonly used for biometric recognition (e.g., face identification). However, these embeddings, or templates, can inadvertently expose sensitive attributes such as age, gender, and ethnicity. Leaking such information can compromise personal privacy and affect civil liberty and human rights. To address these concerns, we introduce a multi-layer protection framework for embeddings. It consists of a sequence of operations: (a) encrypting embeddings using Fully Homomorphic Encryption (FHE), and (b) hashing them using irreversible feature manifold hashing. Unlike conventional encryption methods, FHE enables computations directly on encrypted data, allowing downstream analytics while maintaining strong privacy guarantees. To reduce the overhead of encrypted processing, we employ embedding compression. Our proposed method shields latent representations of sensitive data from leaking private attributes (such as age and gender) while retaining essential functional capabilities (such as face identification). Extensive experiments on two datasets using two face encoders demonstrate that our approach outperforms several state-of-the-art privacy protection methods.

I. INTRODUCTION

The explosive advancement of deep neural networks (DNNs) has revolutionized computational capabilities across domains, from computer vision [5] and speech processing [31] to medical diagnostics [50]. Architectural innovations and optimization techniques have led to significant performance improvements. Research on the privacy and security of DNNs have gained momentum due to stringent regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) and US executive mandates [35]. These considerations are particularly crucial in domains that process sensitive information like biometric systems. Biometric systems use traits such as face, fingerprints, iris, or voice to automatically recognize individuals. In such fields, the trade-off between model utility and privacy guarantees demands precise quantification and optimization.

The core functionality of DNN models relies on projecting very high-dimensional input data into low-dimensional spaces, commonly known as latent representations or embeddings, to facilitate decision-making processes. Many critical applications, like biometric recognition, employ DNNs to improve recognition accuracy. Face embeddings have found widespread utility in facial analytics, spanning applications
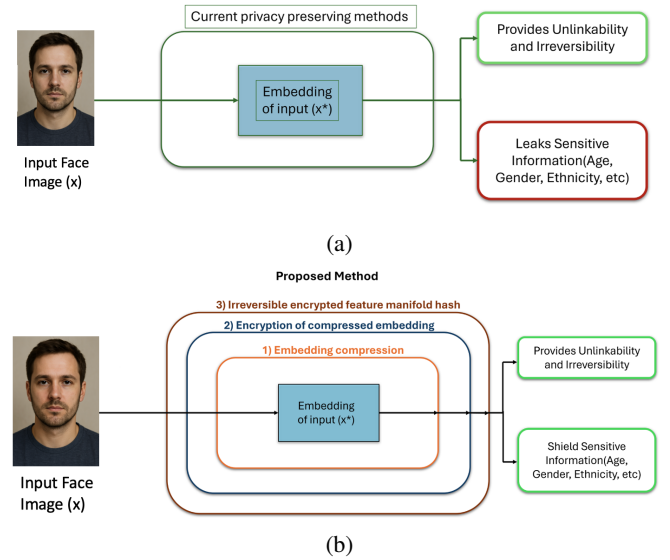
Fig. 1: (a) Existing embedding protection techniques typically guarantee unlinkability and irreversibility but leak sensitive information like age, gender, and ethnicity. (b) Our proposed method improves on the current methods by shielding the sensitive information against privacy attacks while still ensuring unlinkability and irreversibility. The face image used in the figure is AI-generated.

from law enforcement and border security to consumer device authentication [63][1][67]. While many of these applications serve beneficial purposes, the unrestricted deployment of these technologies without adequate privacy safeguards raises significant concerns. The potential for malicious exploitation of facial data presents tangible privacy risks to individuals [13][45], emphasizing the need for privacy-preserving mechanisms in DNN models.

Our research examines the privacy vulnerabilities of embeddings through three distinct attack vectors: Membership Inference, Attribute Inference, and Reconstruction Attacks. The seminal work on membership inference by Shokri et al. [58] established a foundation for understanding data leakage in machine learning models. This led to subsequent investigations into reconstruction attacks, which aim to recover complete or partial training data [37][15]. Parallel research has revealed the vulnerability of models to attribute inference attacks, where unintended characteristics can be extracted from the learned representations [52]. For example, deep

learning models trained for identity verification using face images can inadvertently expose user demographics like ethnicity, gender, and age. While acknowledging the broader landscape of adversarial machine learning, including adversarial examples and data poisoning, this work specifically addresses two fundamental privacy questions: the quantifiable extent of sensitive information extractable from embeddings, and the development of privacy-preserving mechanisms that maintain model utility.

To address the identified embedding vulnerabilities, we examine well-known privacy-preserving approaches. Differential Privacy techniques introduce controlled noise to combat property inference attacks [49], providing theoretical privacy guarantees. Another commonly employed technique is cancelable biometrics in authentication systems [64], where biometric data are encoded with an application-specific distortion scheme. Traditional encryption schemes effectively protect data at rest and during transmission [19][27], but they prohibit computational operations on encrypted data. Fully Homomorphic Encryption (FHE) [47][30] addresses this limitation by enabling direct computation on encrypted embeddings.

Contributions: Our research systematically evaluates several privacy protection techniques, focusing specifically on Inversion Attacks (Membership Inference and Reconstruction) and Attribute Inference Attacks on embeddings. We conduct our analysis using a mix of face embedding models - AdaFace [26] and ArcFace [11] - while examining the leakage of soft biometric attributes (age, gender, ethnicity). Our key contributions are as follows: (1) We establish that Differential Privacy with Laplacian noise successfully protects soft biometric attributes, but at the cost of degraded primary task performance in face recognition; (2) We examine how irreversible feature manifold hashes contribute to privacy in face embeddings. We explore Template Protection schemes as some irreversible feature manifold hashes. Although they effectively counter Inversion Attacks, they remain vulnerable to Attribute Inference Attacks. We validate these findings across three protection schemes available in the research literature: PolyProtect (PP), Negative Face Recognition (NFR), and Minimum Information Units (MIU) [18] [60][61]. (3) We introduce a novel multi-layer impenetrable shield (Fig. 1 (b)) that achieves optimal privacy-utility balance. Our approach combines Fully Homomorphic Encryption (FHE) and an irreversible feature manifold hash on a compressed embedding to significantly enhance privacy while minimally impacting primary task performance.

II. THREAT MODEL

We define *image analytics* as the process of extracting semantic information from an image, which may include sensitive attributes such as age, gender, or ethnicity — commonly known as soft biometrics [55], [65]. The potential to infer these soft biometric cues from facial images, or even from their embeddings, through automated techniques, poses privacy concerns [10]. Such techniques, often implemented using machine learning models like SVMs or DNNs, can enable the extraction of personal information. For instance, if a biometric image (or its embedding) is compromised, it may allow unauthorized parties to derive sensitive biometric details, exposing private information of an individual.

While there are challenges in the training phase of the deep model, we focus on the leakage of side information during the inference stage. Consider a user who submits their face image or embedding to a cloud service to perform face recognition; the service provider can now learn several other details of the user beyond performing the main task. In our work, we presume that the embedding is provided in an encrypted form and the server has no access to the secret key to decrypt it. Our goal is to ensure that the encrypted embedding does not reveal any soft biometric information to unauthorized users. Note that the threat remains unchanged even if the parameters of the models used to extract soft biometric information (e.g., weights of a DNN) are encrypted.

We consider the most challenging threat model according to ISO/IEC 30316, which is the full disclosure model, where the attacker possesses complete knowledge of the irreversible feature manifold hash method for the embedding, including its algorithm, user-specific parameters, and one or more hashes corresponding to an embedding. In addition, we assume that the public key used in FHE is available, but not the private key. If the embeddings are *not* encrypted, the hacker can infer soft biometric attributes from the irreversible hash of the embeddings, as shown in Table III.

A. FHE Basics

Encryption is the process by which plain-text data is encrypted into ciphertext using a secret key and a cryptographic algorithm. Only authorized entities with a private key can decrypt the ciphertext back to the plaintext. Encryption is essential for securing sensitive data from unauthorized access or modification. Homomorphic encryption (HE) is a cryptographic system that permits certain computations to be performed on encrypted data without requiring decryption. In this system, we have public (pk) and private secret (sk) keys, encryption (E) and decryption (D) mechanisms, and plaintext values x and y. When x and y are encrypted as $x' = E(x, pk)$ and $y' = E(y, pk)$, respectively, a cryptosystem is considered homomorphic with respect to a chosen operator (e.g., addition or multiplication), denoted as , if we can find another operator such that $xy = D(x'y', sk)$. This means that we can conduct operations on encrypted data and obtain the same result when decrypting using the private secret key.

Specifically, given $c_i = \mathcal{E}(x_i, pk), i = 1, 2, \cdots, K$, an FHE scheme allows the computation of $c = g(c_1, c_2, \cdots, c_K)$ such that $\mathcal{D}(c, sk) = f(x_1, x_2, \cdots, x_K)$ for any arbitrary function f.

It is essential to note that three types of homomorphic encryption exist [43]: (i) Partial Homomorphic Encryption (PHE) permits addition or multiplication operations; (ii) Somewhat Homomorphic Encryption (SHE) allows limited computations on ciphertexts; (iii) Fully Homomorphic Encryption (FHE) enables computations on ciphertexts of any depth and complexity.

Numerous FHE systems have been introduced, among them are the "Brakerski/Fan-Vercauteren" (BFV), "Brakerski, Gentry, Vaikuntanathan" (BGV), and "Cheon, Kim, Kim and Song" (CKKS) schemes [17]. The BFV and BGV schemes enable vector operations involving integers, while the CKKS scheme facilitates floating-point operations. These schemes achieve Single Instruction Multiple Data (SIMD) operations by bundling plaintext values into an array and then encrypting them to get ciphertext. In this work, we use the HEAAN [8] library based on the CKKS scheme for FHE computations.

III. PRIOR WORK

We investigate a range of privacy-enhancing technologies in biometrics to safeguard sensitive user data.

Privacy in Facial Analytics. A significant body of research investigates methods for enhancing privacy of soft biometrics in face analytics at both the image and embedding levels. PFRNet [6], SensitiveNets [42], PrivacyNet [41], and Multi-IVE [40], each employ distinct approaches such as Autoencoder framework, adversarial regularization, and Incremental Variable Elimination to suppress sensitive attribute information in face embeddings. Unlike our proposed method, these approaches do not fully suppress soft biometrics leakage to the desired level of random guessing. Some work resort to disentangled representation learning as a privacy-preserving mechanism [22][56]. However, disentanglement does not guarantee irreversibility and unlinkability.

Embedding Compression. Compressing embeddings is another useful method for privacy enhancement, especially for deep embedding-based systems. Deep embeddings are often high-dimensional vectors that can contain privacy-sensitive information. Since compression can induce the loss of unwanted information, other sensitive information will likely become difficult to extract. Assine et al. [2] propose a PCA-based compressing embedding representation. In [32], a disentangling mechanism for graph embeddings is proposed to protect users' private information in social media data. [33] present an adaptive bit allocation scheme which assigns different bits to dimensions in the embedding vector based on the importance of a dimension.

Based on their importance, Zhao et al. [33] present an adaptive bit allocation scheme to assign different bits to different dimensions of embedding vectors. In our work, we use Matryoshka Representation Learning (MRL) [28] optimized for face recognition to suppress sensitive attributes while maintaining recognition performance.

Differential Privacy. Differential privacy [49] is often used for enhancing privacy in many application areas, including medical data [12], social network analysis [24], online reviews [23], and face recognition [7][38] [48]. In [66], a differential privacy mechanism is proposed to protect against inversion attacks from confidence scores. However, to our knowledge, no work has been reported that addresses the protection of ancillary data or the privacy of face embeddings.

FHE for Facial Analytics and Medical Imaging. Several papers [46], [47] have shown that existing soft biometric protection approaches are susceptible to reversibility and leakage. In our research, we demonstrate the efficacy of FHE in restricting soft biometric leakage in face embeddings to levels equal to or lower than random guessing and providing a more stringent theoretical guarantee of irreversibility compared to existing methods. FHE has been used in prior work with respect to face recognition. Several papers [44][51][14][25] propose various efficient approaches to perform face matching in the FHE domain using encrypted face embeddings for secure face authentication. In [47][30], the authors show the susceptibility of homomorphic encryption to leaking soft biometric attributes when the private key is not secured or when certain statistical operations such as mean and variance are computed. We address this by employing an additional layer of protection - a non-linear and non-invertible feature transform - thereby providing a dual-layer security. Recent works have proposed methods to make deep learning computations more efficient in the FHE domain, like [59], where the authors propose an encryption-friendly knowledge distillation method to create a framework for confidential and private decentralized training. In [29], the authors propose an encrypted transfer learning technique with efficient matrix multiplication and precise SoftMax approximation in the FHE domain. In [20], the authors propose to improve secure outsourcing computation in the IoT environment of medical devices using FHE, by optimizing the communication and computation burden in the IoT environment. In [21], the authors develop a CNN using an FHE library to perform inference on encrypted data.

Hybrid Protection Techniques. A few papers in the literature have attempted to combine multiple privacy-preserving techniques to provide hybrid protection. In [57], the authors propose to combine cancelable biometric methods like bio hashing, MLP and index of maximum, along with FHE, for hybrid protection. In [4], the authors utilize homomorphic trans-ciphering to improve the security of FHE systems against offline decryption attacks, while [3] uses a combination of bloom filters and homomorphic encryption for template protection.

IV. PROPOSED APPROACH

First, we demonstrate that the embeddings computed from face images leak soft biometric information. For that purpose, we use two well-known face encoders, AdaFace [26] and ArcFace [11], examining the leakage of soft biometric attributes (age, gender, ethnicity) via their embeddings. While embedding protection methods (also known as template protection) have been proposed in the biometric literature, we show that several such non-linear and non-invertible feature transforms do not protect soft biometric attributes. Three such irreversible feature manifold hashes from the literature are used to demonstrate this.

To address the pressing need for an effective privacy protection method that maintains a balance between privacy (of attributes) and utility (of face recognition), we introduce a

novel robust shield centered around FHE-based computations on encrypted embeddings. Unlike conventional encryption methods, FHE supports computations on encrypted data, albeit with high computational overhead. To mitigate this, we incorporate embedding compression using Matryoshka Representation Learning (MRL), which retains core functionality in a reduced representation, thus enhancing computational efficiency within the encrypted domain. These compressed embeddings are subsequently encrypted via FHE, followed by the application of an irreversible feature manifold hash for additional security. The embedding hash ensures that the embedding will be protected by at least 1 layer of security in case of secret key leaks. In our study, we implement PolyProtect [18] as the embedding hash and provide empirical evaluations to confirm the effectiveness of our framework.

V. EXPERIMENTS

This section details experiments broadly based on the protection techniques described above. The goal in each of these experiments is to demonstrate the possibility of soft biometric leakage from face embeddings.

A. Datasets

We conducted our experiments using *CelebSet* [53], [54] and *Chicago Face Database (CFD)* [36], with dataset statistics presented in Table I and Table II, respectively. As the CFD dataset does not provide age information, we employed the 'AgeRated' float values, rounding them to the nearest integer for consistency.

B. Irreversible Feature Manifold Hash techniques

In our study, we examine three template protection methods (embedding hash techniques) proposed in the literature, demonstrating that all of them still inadvertently leak soft biometric information. 1) PolyProtect (PP) [18] transforms a face embedding to a more secure template, using a mapping based on multivariate polynomials parameterized by user-specific coefficients and exponents. 2) Negative Face Recognition (NFR) [60] method produces two templates for each face, positive and negative, where the latter is privacy-protected. Face embeddings are enlarged to a size of 4096 and transformed to obtain a positive template. The negative template, which is the protected reference template, is obtained by randomly replacing each value in the positive template with any of the other values within the vector. 3) Minimum Information Units (MIU) [61] partitions the face embedding into blocks of size k, and these blocks are shuffled to obtain a protected embedding. Table III shows the soft biometric leakage from embeddings protected by irreversible feature manifold hashes.

To validate our proposed multi-layer impregnable shield, we adopt PolyProtect (PP) [18] across all experiments. PP operates with three primary parameters — 'm', 'C', and 'overlap' — and following our ablation study in Fig. 2, we set $m = 5$, $C = 50$, and $overlap = 0$ to maximize protection. We hypothesize that template protection schemes, which operate in the embedding space, can be applied to a variety of embeddings to ensure irreversibility and unlinkability [18].

C. Embedding Compression

We explore embedding compression through **Matryoshka Representation Learning (MRL)** [28] as a technique to improve both privacy in the plaintext domain and processing speeds in the encrypted domain. Based on the ablation study in Fig. 3, we can infer that a compressed embedding size of 64 retains the best performance across all attributes. We train an MRL model to compress embeddings while computing the loss based on identity. Subsequently, we train classifiers to extract soft biometric attributes and predict identity based on the compressed embeddings.

D. Differential Privacy

We investigate the efficacy of Differential Privacy in mitigating soft biometric leakage while maintaining biometric recognition accuracy. To introduce Differential Privacy into our embeddings, we employ the Laplacian mechanism governed by two parameters - sensitivity (Δf) and privacy budget (ϵ) - as outlined in [16][9]. Since the maximum distance between two normalized embeddings is utmost 2, we set the sensitivity parameter to 2. From the ablation study in Fig. 4, we can infer that a privacy budget $\leq 10^{-1}$ provides maximum privacy. A noise vector of the same size as the embeddings is generated, composed of values drawn from the Laplacian distribution. Subsequently, we add this noise vector to the embeddings to form a DP-protected embedding. All downstream tasks are conducted with the DP-protected embeddings.

E. Fully Homomorphic Encryption (FHE)

FHE is an encryption technique that allows computations on encrypted data. The permitted computations are restricted to additive and multiplicative operations only, thereby posing a significant challenge to conduct non-linear operations on ciphertexts. In this paper, we adapt all non-linear functions, such as activation functions, cosine similarity computations, and PolyProtect (PP) [18], in innovative ways to operate in the encrypted domain. We utilize the HEAAN library [8] to perform all FHE experiments.

PolyProtect in FHE. PolyProtect (PP) is a multivariate polynomial function designed to operate in the plaintext domain. We adapt it to FHE by using Taylor series approximations and packing m values of the embedding in a single ciphertext. Subsequently, we obtain a vector of ciphertexts, which are then accumulated with the appropriate ciphertext rotations.

Primary task. In our experiments, we focus on face identification as the primary task on the two datasets. We use the cosine distance metric to compute the match score between two *encrypted* (ciphertext) embeddings and a trained multi-layer perceptron to compute the match score between two *unencrypted* (plaintext) embeddings. First, embeddings from the training set undergo MRL+FHE+PP processing and are then stored as encrypted embeddings. Next, we compute

Gender		Age				Ethnicity			
Males	**Females**	**3-22**	**23-40**	**41-59**	**60-76**	**Hispanic**	**White**	**Black**	**Asian**
38080	34409	5279	43357	22781	1072	738	57873	13414	464
52.50%	47.50%	7.28%	59.82%	31.43%	1.47%	1.01%	79.85%	18.50%	0.64%

TABLE I: Statistics of the CelebSet dataset (80 Identities).

Gender		Age (Rounded to the nearest integer)				Ethnicity			
Males	**Females**	**17-24**	**25-27**	**28-32**	**33-56**	**Asian**	**Black**	**Latino**	**White**
290	307	177	148	143	129	109	197	108	183
48.57%	51.42%	29.64%	24.79%	23.95%	29.61%	18.26%	32.99%	18.09%	30.65%

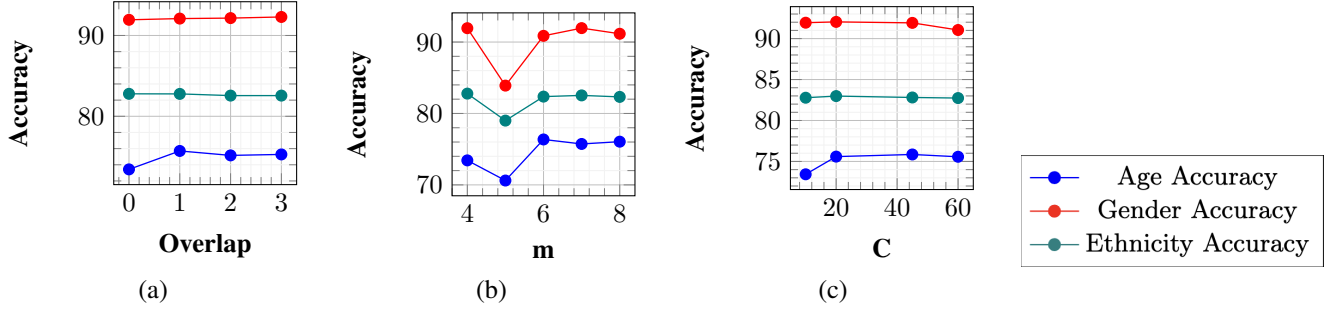TABLE II: Statistics of the Chicago Face Database (597 Identities).

Fig. 2: Ablation Study with PolyProtect shows soft biometric leakage under different settings. (a) Overlap (b) m - Length of polynomial coefficients/exponents (c) [-C,C] - Range of polynomial coefficients.

Model	Dataset	Protection technique	Id (%)	Classification accuracy (↓)		
				Gen (%)	**Age (%)**	**Ethn (%)**
ArcFace	CelebSet	PP	84.40	85.60	66.90	77.87
		MIU	99.68	84.06	66.70	89.02
		NFR	99.69	84.60	67.50	66.13
	CFD	PP	85.04	74.10	81.66	58.33
		MIU	84.41	71.90	83.34	51.50
		NFR	85.21	75.00	87.50	49.16
AdaFace	CelebSet	PP	86.31	93.93	71,35	89.12
		MIU	91.37	88.20	69.70	79.40
		NFR	83.43	88.68	70.13	69.18
	CFD	PP	88.46	80.80	82.50	66.60
		MIU	86.51	74.16	86.60	52.50
		NFR	87.22	83.33	87.50	58.30

TABLE III: Soft biometric leakage in terms of attribute classification accuracy (%) using various Template Protection schemes (CFD - Chicago Face Dataset; Gen - Gender; Ethn - Ethnicity; MIU - Minimum Information Units; NFR - Negative Face Recognition; PP - PolyProtect).

the cosine distance between the encrypted test embeddings and the stored embeddings. Since cosine distance involves an inverse-square root calculation (making it a non-linear operation), we approximate this inverse square root function as an 8th-degree polynomial through polynomial regression to enable cosine distance computation in the encrypted domain.

Soft biometric attribute classification. To evaluate the classification accuracy of soft biometric attributes from encrypted embeddings, we train SVM classifiers on the ASCII values of the encrypted embeddings.
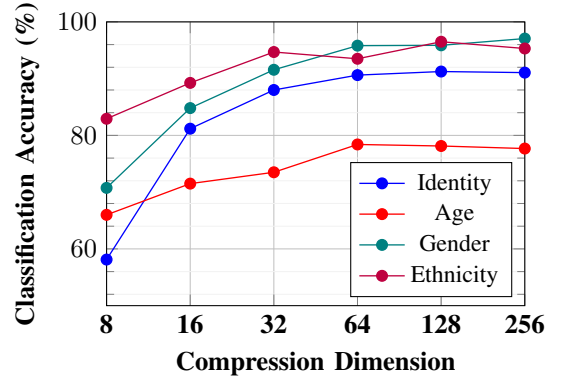

Fig. 3: Performance of attribute extraction (Age, Gender, Ethnicity), along with Rank-1 identification accuracy (Identity), after compression of embeddings using Matryoshka Representation Learning (MRL), as a function of different compression dimensions. The graph is based on an experiment performed using AdaFace on the CelebSet dataset.

F. Execution Times

The execution times for all experiments are presented in Table IV. Since HEAAN [8] supports SIMD operations on ciphertext, we did not conduct separate experiments for each encoder. All plaintext domain experiments were performed using Python on a single A100 Nvidia GPU, while the FHE experiments were conducted with the HEAAN library [8], which runs on C++ and requires a Linux OS. It is important to note that the execution time for "MRL + PP" experiment
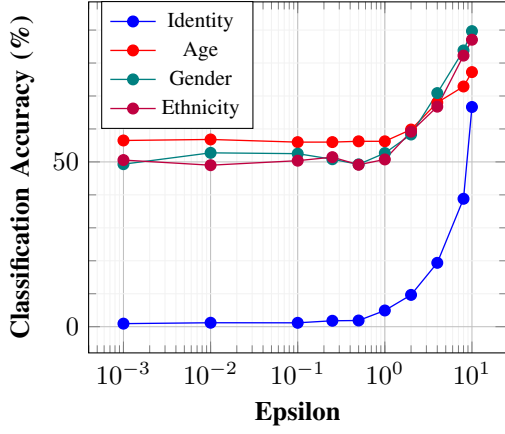
Fig. 4: Classification accuracy of attributes (Age, Gender, Ethnicity) along with Rank-1 identification accuracy (Identity) of face embeddings protected by Differential Privacy as a function of Privacy Budget (ϵ) and Sensitivity ($\Delta f = 2$). The graph is based on an experiment performed using AdaFace on the CelebSet dataset.

Experiment	Execution Time (ms)
MRL (Plaintext domain)	0.5
PP (Plaintext domain)	32
MRL + PP (Plaintext domain)	13
FHE Encryption	85
FHE Decryption	45
FHE + PP	5200
MRL (Plaintext domain) + FHE	12
MRL (Plaintext domain) + FHE + PP	715
Cosine Similarity of 512-sized embeddings (Identification)	9478
Cosine Similarity of 64-sized embeddings (Identification)	1077

TABLE IV: We report the execution times of the proposed method — MRL+FHE+PP. Note that the execution time for MRL+PP does not simply equal the sum of the execution times of MRL and PP due to the embedding size compression in the MRL+PP approach.

does not simply equal the sum of the execution times of MRL and PP experiments; this is because PP is applied to the compressed embedding in the "MRL + PP" experiment. From Table IV, we observe that compressing the embedding (MRL+FHE+PP) provides us with approximately 627% speedup compared to the experiment without compression (FHE+PP). Although the execution times are considerably high, we assert that privacy should not be compromised for speed.

VI. RESULTS

To evaluate our framework for leakage, we have chosen two metrics: Suppression Rate (SR) and Privacy Gain (PG) [39][34][62]. Privacy Gain is defined as PG = $(1 - R_p) - (1 - R_o)$, where, R_o and R_p represent the recognition

performances on the original data and the privacy-enhanced data, respectively. A positive value of PG signifies enhanced data protection. The ideal value for PG is 1. Suppression Rate measures the difference in attribute prediction accuracy with and without privacy enhancement, A_p and A_o, respectively: SR = $(A_o - A_p)/(A_o)$.

As evident in Tables V, VI, Template Protection, Embedding Compression, and Differential Privacy, all leak a substantial amount of soft biometric information. On the other hand, our proposed approach of employing embedding compression through MRL, encryption using FHE, and template protection using PolyProtect demonstrates a high Privacy Gain. This shows that it prevents soft biometric leakage with minimal loss in face recognition accuracy. Fig 5 plots the FMR vs FMNR curve for face verification, where a smaller Area-Under-Curve indicates better performance. We can also observe that our approach reduced the classification accuracies of soft biometric attributes to the level of random chance in both datasets. We could also achieve an almost ideal Privacy Gain in certain scenarios, but we believe this could be due to the imbalanced nature of our datasets.
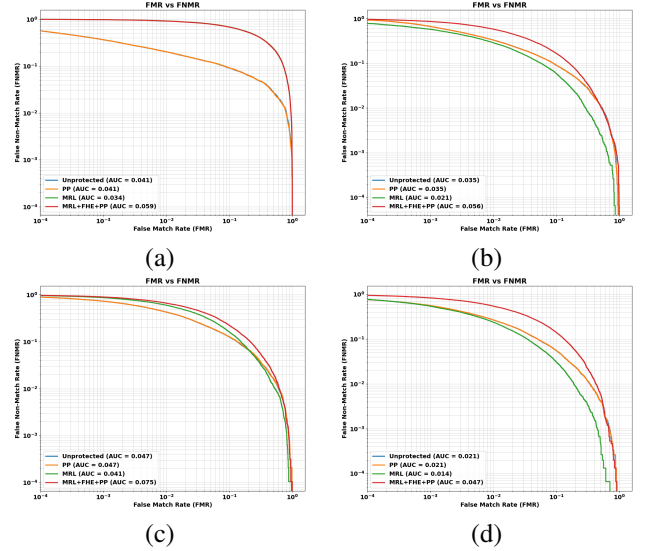

Fig. 5: False Match Rate (FMR) vs False Non-Match Rate (FNMR) for identity verification performance using unprotected and privacy-protected embeddings - (a) AdaFace - CFD; (b) AdaFace - CelebSet; (c) ArcFace - CFD; (d) ArcFace - CelebSet. A smaller area-under-curve signifies better verification performance. PP curve overlaps the "unprotected embeddings" curve, showing a negligible loss in identity verification performance.

From Table VII, we observe soft biometric leakage in all scenarios except when embeddings are encrypted. In cases where embeddings are protected through homomorphic encryption, the AUC score remains below 0.55 across all instances, indicating that the classification performance is reduced to near-random levels. Since gender and ethnicity classification problems are multi-class, we have used macro-averaging to compute the AUC score.

Model	Dataset	Protection technique	Id (%)	Classification accuracy (↓)		
				Gen (%)	Age (%)	Ethn (%)
ArcFace	CelebSet	None	99.42	98.12	87.68	98.81
		PP	99.42	97.35	85.00	98.06
		MRL	97.93	98.00	85.87	97.38
		DP	1.28	49.49	59.23	53.37
		MRL+PP	97.00	96.32	87.32	98.44
		MRL+FHE	97.83	**52.22**	**6.12**	**8.01**
		MRL+FHE+PP	96.95	**52.22**	**6.12**	8.03
	CFD	None	85.06	95.25	94.40	91.97
		PP	85.04	95.07	93.97	91.62
		MRL	84.42	92.02	93.98	91.65
		DP	0.20	51.23	58.65	77.81
		MRL+PP	84.31	90.00	87.90	87.70
		MRL+FHE	84.38	49.98	28.00	24.01
		MRL+FHE+PP	84.28	**49.50**	**27.82**	**24.00**
AdaFace	CelebSet	None	99.41	96.93	90.25	96.31
		PP	99.38	95.75	90.18	96.56
		MRL	97.95	97.63	85.50	98.94
		DP	1.32	48.12	59.17	54.66
		MRL+PP	97.59	96.82	89.13	98.25
		MRL+FHE	97.91	**52.22**	6.12	8.02
		MRL+FHE+PP	97.55	**52.22**	**6.11**	**8.01**
	CFD	None	88.55	94.97	95.02	91.72
		PP	88.46	94.07	94.62	89.62
		MRL	88.33	92.55	93.83	87.43
		DP	0.17	50.98	58.44	80.88
		MRL+PP	88.29	90.58	88.66	85.58
		MRL+FHE	88.23	49.98	27.97	24.02
		MRL+FHE+PP	88.21	**49.50**	**27.90**	**24.00**

TABLE V: Identification accuracy and soft biometric classification accuracy using different protection techniques (DP - Differential Privacy; FHE - Fully Homomorphic Encryption; MRL - Matryoksha Representation Learning; PP - PolyProtect). The highest score has been denoted in **bold**.

Model	Dataset	Protection technique	Id(%)	Privacy Gain (↑)			Suppression Rate (↑)		
				Gender	Age	Ethn	Gender	Age	Ethn
ArcFace	CelebSet	PP	99.42	0.72	2.68	0.75	0.0073	0.0306	0.0075
		MRL	97.93	0.12	1.81	1.43	0.0012	0.0206	0.0145
		DP	1.28	**48.63**	28.45	45.44	**0.4956**	0.3244	0.4590
		MRL+PP	97.00	1.80	0.36	0.37	0.0183	0.0041	0.0037
		MRL+FHE	97.83	45.90	**81.56**	**90.80**	0.4678	**0.9302**	**0.9189**
		MRL+FHE+PP	96.95	45.90	**81.56**	90.78	0.4678	**0.9302**	0.9187
	CFD	PP	85.04	0.18	0.43	0.35	0.0019	0.0046	0.0038
		MRL	84.42	3.23	0.42	0.32	0.0339	0.0044	0.0035
		DP	0.20	44.02	35.75	14.16	0.4600	0.3700	0.1500
		MRL+PP	84.31	5.25	6.50	4.27	0.0551	0.0689	0.0464
		MRL+FHE	84.38	45.27	66.40	67.96	0.4753	0.7034	0.7389
		MRL+FHE+PP	84.28	**45.75**	**66.58**	**67.97**	**0.4803**	**0.7053**	**0.7390**
AdaFace	CelebSet	PP	99.38	1.18	0.07	-0.25	0.0122	0.0008	-0.0026
		MRL	97.95	-0.70	4.75	-2.63	-0.0072	0.0526	-0.0273
		DP	1.32	**48.81**	31.08	41.65	**0.5000**	0.3400	0.4300
		MRL+PP	97.59	0.11	1.12	-1.94	0.0011	0.0124	-0.0201
		MRL+FHE	97.91	44.71	84.13	88.29	0.4613	0.9322	0.9167
		MRL+FHE+PP	97.55	44.71	**84.14**	**88.30**	0.4613	**0.9323**	**0.9168**
	CFD	PP	88.46	0.90	0.40	2.10	0.0090	0.0040	0.0220
		MRL	88.33	2.42	1.19	4.29	0.0250	0.0120	0.0460
		DP	0.17	**43.99**	36.58	10.84	**0.4632**	0.3840	0.1100
		MRL+PP	88.29	4.39	6.36	6.14	0.0460	0.0660	0.0670
		MRL+FHE	88.23	39.99	57.05	57.70	0.4444	0.6710	0.7061
		MRL+FHE+PP	88.21	40.47	**57.12**	**57.72**	0.4498	**0.6718**	**0.7063**

TABLE VI: Privacy Gain and Suppression Rate across different soft biometric attributes using different protection techniques (DP - Differential Privacy; FHE - Fully Homomorphic Encryption; MRL - Matryoksha Representation Learning; PP - PolyProtect). The highest score has been denoted in **bold**.

Model	Dataset	Protection Technique	Age (↓)	Gen (↓)	Ethn (↓)
ArcFace	CelebSet	None	0.88	0.99	0.99
		PP	0.98	0.99	0.99
		MRL	0.88	0.99	0.99
		DP	0.52	**0.47**	**0.50**
		MRL+PP	0.97	0.99	0.98
		MRL+FHE	**0.46**	0.50	0.52
		MRL+FHE+PP	**0.46**	0.50	0.52
	CFD	None	0.98	0.99	0.98
		PP	0.99	0.99	0.98
		MRL	0.98	0.99	0.99
		DP	**0.50**	**0.49**	0.51
		MRL+PP	0.99	0.99	0.98
		MRL+FHE	**0.50**	0.50	0.49
		MRL+FHE+PP	0.51	**0.49**	**0.47**
AdaFace	CelebSet	None	0.84	0.99	0.99
		PP	0.88	0.99	0.99
		MRL	0.88	0.99	0.99
		DP	**0.49**	**0.48**	0.51
		MRL+PP	0.90	0.99	0.99
		MRL+FHE	0.50	0.52	**0.49**
		MRL+FHE+PP	**0.49**	0.51	0.50
	CFD	None	0.92	0.99	0.97
		PP	0.98	0.99	0.99
		MRL	0.94	0.99	0.98
		DP	0.49	0.52	0.50
		MRL+FHE	**0.48**	0.51	**0.47**
		MRL+FHE+PP	**0.48**	**0.49**	0.50

TABLE VII: Area-Under-Curve (AUC) scores for soft biometric classification using various protection techniques (DP - Differential Privacy; FHE - Fully Homomorphic Encryption; MRL - Matryoksha Representation Learning; PP - PolyProtect). The lowest score has been denoted in **bold**
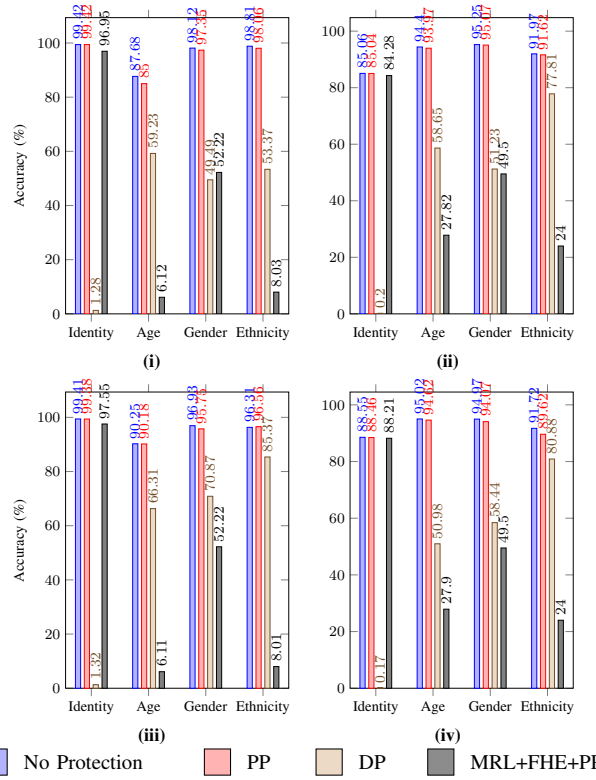

Fig. 6: Identification accuracy (Identity) and attribute classification accuracy (Age, Gender, Ethnicity) across different privacy techniques: (i) ArcFace on CelebSet; (ii) ArcFace on CFD; (iii) AdaFace on CelebSet; (iv) AdaFace on CFD.

VII. ANALYSIS

Why is the combination of FHE+PP not considered? MRL's contribution to privacy gain is negligible, but, it contributes significantly in improving processing time. From Table IV, we can observe a 627% speedup of our framework with embedding compression. The execution times for all FHE combinations are shown in Table IV.

How does our framework evaluate against DP? DP provides maximum protection to our embeddings, as shown in Table V - the accuracy values indicate randomness. However, our application requires us to conserve the primary functionality of embeddings (i.e., face identification or verification) while protecting soft biometric attributes. With DP, the embeddings lose their primary functionality. On the other hand, our proposed framework retains the primary task of the embeddings.

Why is the order MRL+FHE+PP important? MRL is applied first to reduce the dimensionality of the embedding, thereby improving the computational efficiency of FHE. We choose to apply PP on FHE-encrypted embedding for privacy reasons. In cases where multiple PP embeddings and user parameters are leaked to the attacker, the likelihood of inversion increases [18]. On the other hand, when the embeddings are encrypted before the application of PP, it would not compromise privacy.

VIII. CONCLUSION

In this study, we investigated various protection techniques to secure face embeddings from potential attacks and privacy breaches. Our experimental results indicate that no single approach — whether Template Protection, Differential Privacy, or Fully Homomorphic Encryption — is sufficient to independently safeguard embeddings effectively. To address this, we propose a robust, multi-layered protection strategy combining (1) Embedding Compression, (2) Fully Homomorphic Encryption, and (3) Irreversible Feature Manifold Hash applied within the FHE space. This integrated approach achieves optimal embedding security by shielding soft biometric attributes while maintaining high biometric recognition accuracy, as validated through comprehensive experiments across two datasets and two distinct face encoders. By leveraging FHE's theoretical guarantees, this strategy ensures that only authorized individuals with the correct secret key can access computation results. Moreover, even when the secret key is compromised, the irreversible feature transform serves as an additional safeguard to prevent full exposure of the embeddings.

IX. ETHICAL IMPACT STATEMENT

The overall approach in our research is focused on protecting leakage of privacy in facial analytics. Publicly available face datasets (CelebSet [53][54] and Chicago Face Database [36]) have been used for reporting accuracy and performance evaluation. We do not show any of the dataset images / personal data in the paper. Our research does not involve data collection from human subjects. The contributions do

not include the creation of a new dataset. To the best of our knowledge, we do not see any potential negative implications on the society from our paper.

REFERENCES

[1] A. Alfalou, C. Brosseau, and W. Kaddah. Optimization of decision making for face recognition based on nonlinear correlation plane. *Optics Communications*, 343:22–27, 2015.

[2] J. S. Assine, A. Godoy, and E. Valle. Compressing representations for embedded deep learning, 2019. https://arxiv.org/abs/1911.10321.

[3] A. Bassit, F. Hahn, R. Veldhuis, and A. Peter. Hybrid biometric template protection: Resolving the agony of choice between bloom filters and homomorphic encryption. *IET biometrics*, 11(5):430–444, 2022.

[4] P. Bauspieß, C.-M. Zok, A. Costache, C. Rathgeb, J. Kolberg, and C. Busch. Mt-pro: Multibiometric template protection based on homomorphic transciphering. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2023.

[5] D. Bharatia and P. Rane. Research on advanced image analysis and it's applications. In *Global Conference for Advancement in Technology (GCAT)*, pages 1–5, 2019.

[6] B. Bortolato, M. Ivanovska, P. Rot, J. Križaj, P. Terhörst, N. Damer, P. Peer, and V. Štruc. Learning privacy-enhancing face representations through feature disentanglement. In *15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 495–502, 2020.

[7] M. Chamikara, P. Bertok, I. Khalil, D. Liu, and S. Camtepe. Privacy preserving face recognition utilizing differential privacy. *Computers & Security*, 97:101951, 2020.

[8] J. H. Cheon, A. Kim, M. Kim, and Y. Song. Homomorphic encryption for arithmetic of approximate numbers. In T. Takagi and T. Peyrin, editors, *Advances in Cryptology – ASIACRYPT*, pages 409–437, Cham, 2017. Springer International Publishing.

[9] W. L. Croft, J.-R. Sack, and W. Shi. Differentially private obfuscation of facial images. In A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 229–249, Cham, 2019. Springer International Publishing.

[10] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2016.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[12] A. Dyda, M. Purcell, S. Curtis, E. Field, P. Pillai, K. Ricardo, H. Weng, J. C. Moore, M. Hewett, G. Williams, and C. L. Lau. Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. *Patterns*, 2(12):100366, 2021.

[13] M. Eichelberg, K. Kleber, and M. Kämmerer. Cybersecurity challenges for pacs and medical imaging. *Academic Radiology*, 27(8):1126–1139, 2020.

[14] J. J. Engelsma, A. K. Jain, and V. N. Boddeti. Hers: Homomorphically encrypted representation search. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):349–360, 2022.

[15] I. Fábián and G. Gulyás. De-anonymizing facial recognition embeddings. *Infocommunications Journal*, 12:50–56, 01 2020.

[16] Google. Secure noise generation for differential privacy, 2019.

[17] S. Gorantala, R. Springer, and B. Gipson. Unlocking the potential of fully homomorphic encryption. *Commun. ACM*, 66(5):72–81, apr 2023.

[18] V. K. Hahn and S. Marcel. Towards protecting face embeddings in mobile face verification scenarios. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):117–134, 2022.

[19] M. K. Hasan, S. Islam, R. Sulaiman, S. Khan, A.-H. A. Hashim, S. Habib, M. Islam, S. Alyahya, M. M. Ahmed, S. Kamil, and M. A. Hassan. Lightweight encryption technique to enhance medical image security on internet of medical things applications. *IEEE Access*, 9:47731–47742, 2021.

[20] J. Huang, Q. Cui, Z. Zhou, K. Yu, C.-N. Yang, and K.-K. R. Choo. Encrypted domain secret medical-image sharing with secure outsourcing computation in iot environment. *IEEE Internet of Things Journal*, 2023.

[21] N. Jain, K. Nandakumar, N. K. Ratha, S. Pankanti, and U. Kumar. Efficient CNN building blocks for encrypted data. In *Association for the Advancement of Artificial Intelligence (AAAI) Workshop*, 2021.

[22] T. Jang and X. Wang. Fades: Fair disentanglement with sensitive relevance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12067–12076, June 2024.

[23] J. Jia and N. Z. Gong. AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 513–529, Baltimore, MD, Aug. 2018. USENIX Association.

[24] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng. Applications of differential privacy in social network analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):108–127, 2023.

[25] A. K. Jindal, I. Shaik, V. Vasudha, S. R. Chalamala, R. Ma, and S. Lodha. Secure and privacy preserving method for biometric template protection using fully homomorphic encryption. In *IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1127–1134, 2020.

[26] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18729–18738, 2022.

[27] Kiran, B. Parameshachari, H. Panduranga, and S. liberata Ullo. Analysis and computation of encryption technique to enhance security of medical images. *IOP Conference Series: Materials Science and Engineering*, 925(1):012028, Sep 2020.

[28] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, and A. Farhadi. Matryoshka representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc., 2022.

[29] S. Lee, G. Lee, J. W. Kim, J. Shin, and M.-K. Lee. Hetal: efficient privacy-preserving transfer learning with homomorphic encryption. In *International Conference on Machine Learning*, pages 19010–19035, 2023.

[30] B. Li and D. Micciancio. On the security of homomorphic encryption on approximate numbers. In *Advances in Cryptology – EUROCRYPT: 40th Annual International Conference on the Theory and Applications of Cryptographic Techniques, October 17–21, Proceedings, Part I*, page 648–677. Springer-Verlag, 2021.

[31] J. Li et al. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

[32] K. Li, G. Luo, Y. Ye, W. Li, S. Ji, and Z. Cai. Adversarial privacy-preserving graph embedding against inference attack. *IEEE Internet of Things Journal*, 8(8):6904–6915, 2021.

[33] S. Liao, J. Chen, Y. Wang, Q. Qiu, and B. Yuan. Embedding compression with isotropic iterative quantization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8336–8343, Apr. 2020.

[34] Y. Liu, X. Zhu, J. Wang, and J. Xiao. A quantitative metric for privacy leakage in federated learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3065–3069, 2021.

[35] T. K. Lively. Facial recognition in the united states: Privacy concerns and legal developments. *Security Technology*, 2021.

[36] D. S. Ma, J. Correll, and B. Wittenbrink. The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behav Res Methods*, 47(4):1122–1135, Dec. 2015.

[37] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain. On the reconstruction of face images from deep face templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1188–1202, 2019.

[38] Y. Mao, J. Feng, F. Xu, and S. Zhong. A privacy-preserving deep learning approach for face recognition with edge computing. In *USENIX Workshop on Hot Topics in Edge Computing*, 2018.

[39] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021.

[40] P. Melzi, H. O. Shahreza, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, J. Fierrez, S. Marcel, and C. Busch. Multi-ive: Privacy enhancement of multiple soft-biometrics in face embeddings. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 323–331, 2023.

[41] V. Mirjalili, S. Raschka, and A. Ross. Privacynet: Semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412, 2020.

[42] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana. Sensitive-Nets: Learning Agnostic Representations with Application to Face Images . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(06):2158–2164, June 2021.

[43] K. Munjal and R. Bhatia. A systematic review of homomorphic encryption and its contributions in healthcare industry. *Complex & Intelligent Systems*, 9:1–28, 2022.

[44] V. Naresh Boddeti. Secure face matching using fully homomorphic encryption. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, page 1–10, 2018.

[45] National Academies of Sciences, Engineering, and Medicine. *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance*. The National Academies Press, Washington, DC, 2024.

[46] D. Osorio-Roig, P. A. Gerlitz, C. Rathgeb, and C. Busch. Reversing deep face embeddings with probable privacy protection. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2023.

[47] D. Osorio-Roig, C. Rathgeb, P. Drozdowski, P. Terhörst, V. Štruc, and C. Busch. An attack on facial soft-biometric privacy enhancement. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2):263–275, 2022.

[48] L. Ou, Y. He, S. Liao, Z. Qin, Y. Hong, D. Zhang, and X. Jia. Faceidp: Face identification differential privacy via dictionary learning neural networks. *IEEE Access*, 11:31829–31841, 2023.

[49] A. E. Ouadrhiri and A. Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE Access*, 10:22359–22380, 2022.

[50] S. Poudel. A study of disease diagnosis using machine learning. *Medical Sciences Forum*, 10(1), 2022.

[51] G. Pradel and C. Mitchell. Privacy-preserving biometric matching using homomorphic encryption. In L. Zhao, N. Kumar, R. Hsu, and Z. Zou, editors, *Proceedings: IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom, Shenyang, China*, Oct. 2021.

[52] J. Pu and Z. Takhirov. Embedding attack project (work report). Technical report, PnS Privacy Innovation Lab, TikTok, 2024.

[53] D. Raji and J. Lee. Celebset, 2019. "https://tinyurl.com/362s54h9", Accessed: 2024-11-13.

[54] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, page 145–151. Association for Computing Machinery, 2020.

[55] A. Ross, S. Banerjee, and A. Chowdhury. Deducing health cues from biometric data. *Computer Vision and Image Understanding*, 221:103438, 2022.

[56] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni. Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision*, pages 746–761. Springer, 2020.

[57] H. O. Shahreza, C. Rathgeb, D. Osorio-Roig, V. K. Hahn, S. Marcel, and C. Busch. Hybrid protection of biometric templates by combining homomorphic encryption and cancelable biometrics. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2022.

[58] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[59] N. Tastan and K. Nandakumar. Capride learning: Confidential and private decentralized learning based on encryption-friendly distillation loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8084–8092, 2023.

[60] P. Terhörst, M. Huber, N. Damer, F. Kirchbuchner, and A. Kuijper. Unsupervised enhancement of soft-biometric privacy with negative face recognition, 2020. https://arxiv.org/abs/2002.09181.

[61] P. Terhörst, K. Riehl, N. Damer, P. Rot, B. Bortolato, F. Kirchbuchner, V. Struc, and A. Kuijper. Pe-miu: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access*, 8:93635–93647, 2020.

[62] I. Wagner and D. Eckhoff. Technical privacy metrics: A systematic survey. *ACM Comput. Surv.*, 51(3), 2018.

[63] Q. Wang, A. Alfalou, and C. Brosseau. New perspectives in face correlation research: a tutorial. *Advances in Optics and Photonics*, 9(1):1–78, Mar 2017.

[64] W. Yang, S. Wang, J. Hu, X. Tao, and Y. Li. Feature extraction and learning approaches for cancellable biometrics: A survey. *CAAI Transactions on Intelligence Technology*, 9(1):4–25, 2024.

[65] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, pages 1–11, 2024.

[66] D. Ye, S. Shen, T. Zhu, B. Liu, and W. Zhou. One parameter defense: Defending against data inference attacks via differential privacy. *IEEE Transactions on Information Forensics and Security*, 17:1466–1480, 2022.

[67] C. Zhao, X. Li, and Y. Cang. Bisecting k-means clustering based face recognition using block-based bag of words model. *Optik - International Journal for Light and Electron Optics*, 126(19):1761–1766, 2015.