University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 9, Issue 2, 2025
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

A Novel Hybrid Approach with Siamese Networks and Domain-Adversarial Training

Naadiya Mirbahar[1]*, Kamlesh Kumar[2], Asif Ali Laghari[1]

[1] Department of Computer Science, Sindh Madressatul Islam University, Pakistan
[2] Department of Software Engineering, Sindh Madressatul Islam University, Pakistan
Email: naadiya.khudabux@usms.edu.pk, kamlesh@smiu.edu.pk, asif.laghari@smiu.edu.pk

Abstract: Balancing data utility and privacy is a persistent challenge in privacy-conscious machine learning, especially when handling personal or biometric data. We propose PES-DAT (Privacy-Enhanced Siamese Network with Domain-Adversarial Training), a framework that generates discriminative, privacy-preserving embeddings for clustering tasks. PES-DAT combines label-guided contrastive representation learning with a domain-adversarial classifier, enforced through a Gradient Reversal Layer (GRL), to produce task-relevant embeddings that are robust against sensitive-attribute leakage. Because contrastive pairs are formed from available class labels, PES-DAT operates in a semi-supervised regime: labels guide representation learning at training time, while the downstream clustering and deployment require no labels. PES-DAT further tunes the privacy–utility trade-off during training through a dynamically scheduled Laplace mechanism, so that the encoder first learns discriminative structure and the differential-privacy guarantee is progressively tightened as representations stabilize. Evaluations on two benchmark datasets—the UCI Human Activity Recognition (HAC) dataset and a campus-behavior dataset—show that PES-DAT achieves up to 10% higher silhouette scores and reduces privacy leakage by up to 30% relative to the AIB and PPGAN baselines. The results demonstrate that PES-DAT is a robust and flexible solution for privacy-preserving representation learning in semi-supervised settings.

Keywords: Domain-Adversarial Neural Networks (DANN); privacy–utility trade-off; contrastive learning; Gradient Reversal Layer (GRL); differential privacy.

I. INTRODUCTION

With the advent of digital technologies, huge volumes of personally identifiable information are constantly received and processed across various sectors, including health care, education, telecommunications, finance to name but a few. Although data-driven models offer many advantages, they also represent a very considerable hazard to privacy — especially when we consider that their inputs often use personal or biometric data. However, classic anonymisation techniques often prove inadequate: the latest research shows that re-identification attacks can also be successfully directed against non-identifiable data collections. A well-known problem which will exacerbate considerably when disparate datasets are combined and analysed. This is especially relevant in the educational level, where there is a growing trend to process aggregated student behavioral data with demographic information for analytics and targeted assistance while risking unintentional disclosure of sensitive information. To tackle these challenges, we propose PES-DAT, a hybrid learning framework that fuses Siamese networks, domain-adversarial training and a dynamically scheduled differential-privacy mechanism to obtain high utility whilst achieving strong privacy guarantees on representation learning [1].

It has been demonstrated in research over the last few months that embeddings coming out of deep learning and other models expose personally identifiable information [2]. Many becomes available ranging from campus-card transactions to smartphone location data and bus-card usage have alarmed us toward with abilities privacy risks [3]. Researchers have attempted to make such data anonymous but removing or obfuscating personally identifiable information (PII) still leaves us vulnerable to privacy violations [4,45].

Despite the anonymization process, the weaknesses in the statistical analysis of behaviors and finances illustrate the risk of de-anonymization, making it a very important issue in educational institutions where confidential information about the students can be unknowingly exposed due to the aggregation of different data types [5]. A recent study [6] has

proved that even anonymized behaviors such as the way one uses a smartphone or makes transactions via a university card can be coupled with machine learning to de-anonymize individuals with extremely high accuracy, showing how fast it is possible to de-anonymize data. [7].

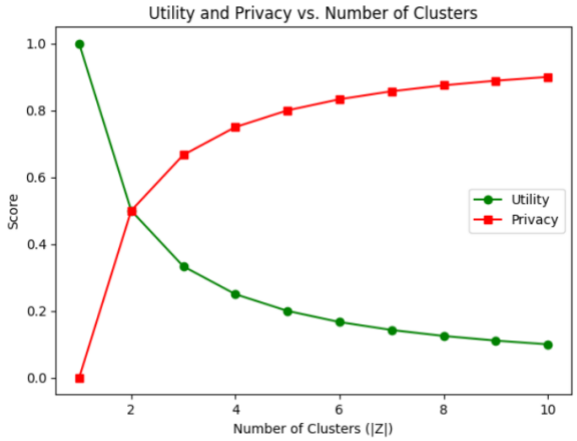

Figure 1. Conceptual illustration of the privacy–utility trade-off

Crucially, the leakage risk is not limited to raw records: the learned representation itself can act as a side channel. Membership-inference attacks can determine whether a particular individual's record was used to train a model from its outputs alone [40], and model-inversion attacks can reconstruct sensitive input attributes from model confidences or embeddings [41]. Because standard representation-learning objectives contain no explicit term to suppress sensitive information, structural embeddings tend to retain attributes such as gender or age even when those attributes are never used as targets. This motivates a representation learner that is explicitly trained to discard sensitive structure while preserving task-relevant structure.

To tackle this challenge, we propose a hybrid approach that couples a Siamese network for learning discriminative representations with adversarial training via a domain classifier, enhanced by a Gradient Reversal Layer (GRL) for domain-invariant learning. We additionally employ a differential-privacy mechanism [8], specifically the Laplace mechanism, to protect sensitive data while preserving model utility [9]. Because the contrastive pairs that drive the Siamese branch are constructed from class labels, PES-DAT is a semi-supervised method: it uses labels to shape the embedding space during training, but the embeddings it produces are clustered without labels at evaluation and deployment time. This design strikes a balance between privacy and utility and offers a robust solution for privacy-preserving representation learning in scenarios where data confidentiality is critical [10].

Contributions of this paper are listed below: (i) a unified architecture for PES-DAT combining label-constrained contrastive learning, domain adversarial training using a GRL and differential privacy in one end-to-end model; (ii) a dynamically scheduleable Laplace mechanism for changing the value of privacy budget during training epochs and thereby enabling the encoder to focus on utility at first and then gradually increase privacy of representations; and (iii) experiments on HAC and campus-behavior datasets to demonstrate better performance with respect to clustering, utility, and privacy when compared to the AIB and PPGAN models.

The rest of the paper is structured as follows. Section 2 surveys relevant literature. Section 3 states the problem and its mathematical formulation. Section 4 introduces the PES-DAT approach with the contrastive Siamese network, domain-adversarial training branch with GRL, and dynamically scheduled Laplace mechanism. Section 5 introduces the PES-DAT algorithm. Section 6 introduces the motivating case study, Section 7 introduces the experimental setup, and Section 8 discusses the results and findings. Section 9 covers future directions, and Section 10 concludes the paper.

II. Related Work

The increasing application of machine learning to privacy-sensitive fields such as health care, finance, education and fashion industry [44] has seen significant research into methods for protecting privacy. Conventional approaches tend to be biased towards utility and performance, which are incompatible with privacy when using sensitive personal and biometric data. The main challenge is thus how to achieve privacy and utility simultaneously – where privacy guarantees that private data cannot be accessed, and utility guarantees that models are accurate and effective [11].

A number of approaches have been suggested to solve the privacy-utility trade-off challenge in machine/deep learning [12–14]. Differential privacy is one of the most popular approaches to privacy protection that provides data security by introducing noise [15]. Even though it is an efficient method, it often degrades the quality of a model, thus contradicting the utility principle, especially for downstream applications [16]. o overcome this, more recently, dynamic privacy budgets that adjust the amount of injected noise depending on the model performance and the sensitivity of data have been introduced [17]. The approach of privacy budget controls the amount of privacy provided in the process of training [18]; for instance, in fraud detection systems [19]

it controls the noise in transaction data [20], Still, an excessive focus on privacy results in inaccurate models, and

Siamese neural networks, which have been developed to classify similar from dissimilar inputs [22], are becoming widely used in privacy-preserving machine learning since they can create discriminative features not using sensitive features explicitly. Siamese networks are often employed for anomaly detection [23], medical imaging comparison [24], and biometric authentication [25]; for instance, for comparing MRI scans with disease progression [26] or detecting financial fraud [27]. By focusing on task-relevant attributes, Siamese networks can learn representations that encode little sensitive information in privacy-critical settings. However, their reliance on stable data distributions can hinder generalization in noisy or dynamic environments [28].

In order to solve such problems, Domain Adversarial Neural Network (DANN) is proposed [29]. DANN extends the encoder by adding an adversarial domain classifier; the adversary attempts to guess a sensitive attribute (gender, age, race), whereas the encoder is trained using gradient reversal layer to hinder the prediction process [30]. The output is then a representation where sensitive attribute is mostly invariant, which makes the learning process more private [31]. DANN has been successfully used in healthcare, education, and wireless networks. Despite their strength in domain-invariant learning, DANNs are difficult to train, and balancing the adversarial loss against other objectives such as clustering quality is challenging and computationally expensive [32, 33].

Contrastive learning, a self-supervised methodology that learns representations by comparing positive and negative sample pairs, has further advanced privacy-preserving methods. Approaches such as SimCLR [34] learn effective representations from unlabeled data, making them attractive when labeled data is scarce. When combined with differential privacy, contrastive learning can help ensure that sensitive attributes are not embedded in the learned representation [35]. A key benefit is the ability to exploit large amounts of unlabeled data [36]; the central difficulty is balancing the privacy–utility trade-off when noise from differential privacy degrades representation quality [37, 38].

III. PROBLEM DEFINITION

In many real-world scenarios, organizations share or publish datasets after removing explicit identifiers such as names, phone numbers, and addresses, assuming that this renders the data anonymous. However, de-identification of data is possible even by quasi-identifiers (such as age, location, or behavior), especially when linked with other data sources, posing privacy risks [39, 40]. Privacy issues also emerge in machine learning algorithms where sensitive features may be indirectly revealed in the process of training a model [35, 41]. The danger is imminent in the case of unsupervised processes such as clustering, wherein the embeddings can end up learning sensitive characteristics even if the labels have been stripped out of them [42]. Standard techniques of representation learning have no means of preventing sensitive features from embedding themselves in the embeddings [43]. The problem that we solve is thus learning embeddings that preserve privacy in the task of clustering. Let Z be the embeddings and λ be the trade-off parameter; the loss function is defined as in Equation (1):

$$max_{f_\vartheta} Utility\ (Z) - \lambda\ .PrivacyLeakage\ (Z) \qquad (1)$$

- Maximizing clustering utility ensures grouping similar instances by task-relevant features for effective clustering.
- Minimizing sensitive-information leakage prevent the model from embedding sensitive attributes that could lead to privacy violations.

We propose PES-DAT, which integrates a Siamese network for learning discriminative embeddings with adversarial domain training to protect sensitive information. This section presents the architecture and training methodology of the Privacy-Enhanced Siamese Network with Domain-Adversarial Training (PES-DAT). The network learns good-quality feature representations for clustering while suppressing sensitive information in the learned embeddings. Figure 2 shows the overall framework.

IV. . METHODOLOGY

This section provides the architecture and training process details of Privacy Enhanced Siamese Network with Domain Adversarial Training (PES-DAT). This network preserves good quality representation of features for clustering purpose while hiding the sensitive features in the learned embeddings. Figure 2 provides an overview of the entire framework.

A. Data Preprocessing

Preprocessing starts with handling missing values where those values are discarded to preserve the integrity of the data set. The two data sets include (1) HAC data set (Human Activity Recognition Using Smartphones), a publicly available data set of smartphone sensor data collected from 30 participants engaged in daily activities; and (2) Campus

data set collected from the university database including behavioral and demographic features. In the campus data set, only those users who exhibit high activity are retained. In

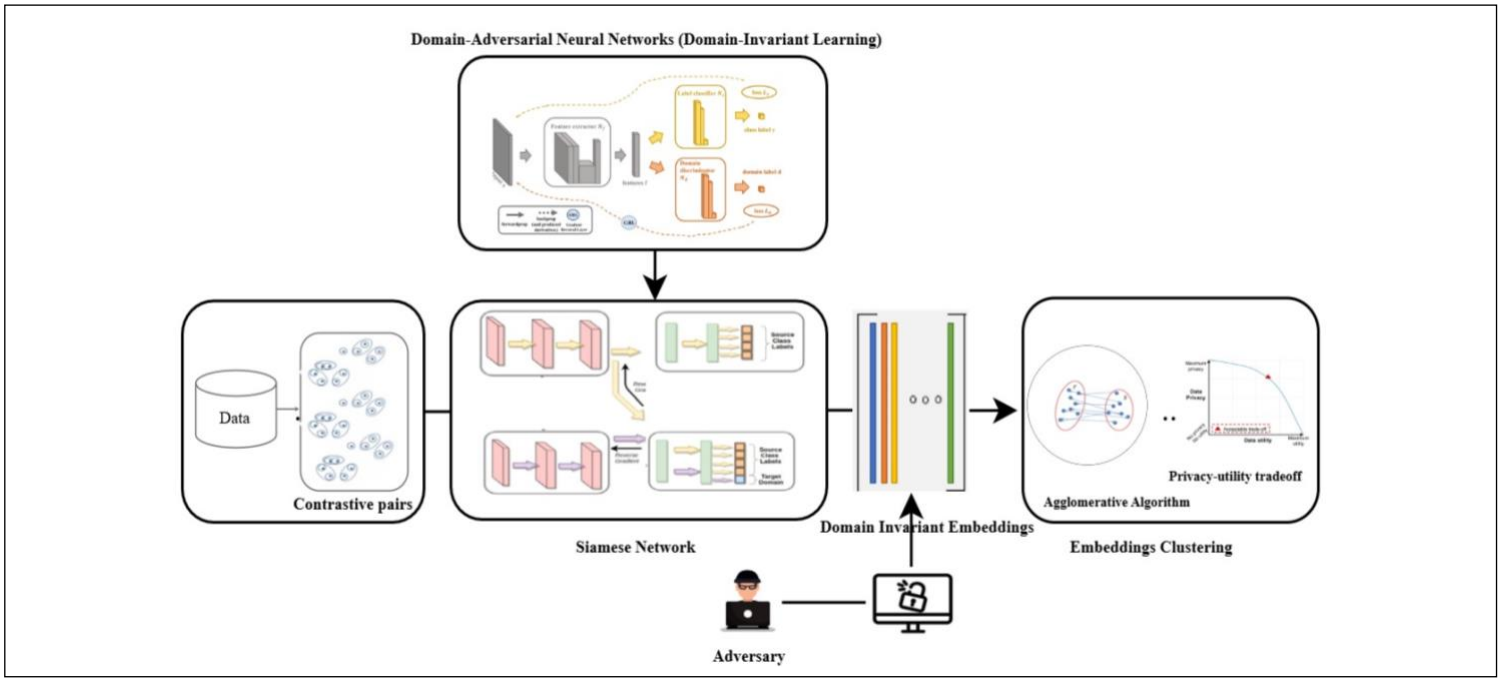

Figure 2. *PES-DAT architecture — Siamese encoder, GRL + adversarial domain classifier, dynamically scheduled Laplace mechanism, and downstream clustering*

order to decrease dimensions and keep important features, SelectKBest using mutual information is employed for feature selection. The selected features are standardized using the StandardScaler in order to make sure that all features have mean 0 and variance 1 such that the contribution of all features is equal.

B. Contrastive Pair Construction

Post processing, contrastive pairs are formed based on the available labels of classes. Positive pairs are chosen from samples which have same labels, say for example, class A (i.e., a pair of class A samples) whereas negative pairs are selected from samples having distinct labels, such as class A and class B. This set of labeled pairs is used to train the Siamese neural network. As the formation of pairs is based on labels, this phase of learning representations can be considered as supervised; thus we can say that PES-DAT is semi-supervised because here the embedding space is formed based on supervision but clusters are formed without any labels during evaluation and deployment.

C. Siamese Networks

Siamese neural network learns the embeddings of the input data. Each branch encodes the input sample into latent vector and contrastive loss tries to pull embeddings of similar pairs close to each other and dissimilar embeddings far away. For an input pair (x_1, x_2) having label y (y=1 if similar and y=0 otherwise), the network will generate embeddings z_1 and z_2. Let the distance between these embeddings be represented as $D = \|z_1 - z_2\|^2$. The contrastive loss is shown in Equation (2):

$$\mathcal{L}_{Contrastive} = \frac{1}{2N}\sum_{i=1}^{N} y_i \left\|z_1^{(i)} - z_2^{(i)}\right\|^2 + (\mathrm{i} - y_i) + \max(0, margin - \left\|z_1^{(i)} - z_2^{(i)}\right\|^2) \qquad (2)$$

Where $y_i \in \{0,1\}$ is the label indicating if the pair x_1, x_2 are similar ($y_i = 1$) or dissimilar ($y_i = 0$). $\left\|z_1^{(i)} - z_2^{(i)}\right\|^2$ is the Euclidean distance between the embeddings and margin is a hyperparameter that sets the minimum distance between dissimilar pairs.

D. Adversarial Network for Domain-Invariant Learning

To enforce domain-invariant, privacy-preserving embeddings, a domain-adversarial branch is attached after the Siamese encoder's embedding layer to prevent sensitive

attributes (e.g., gender, age) from being encoded. A separate adversary network A, parameterized by θ_{adv}, predicts the sensitive attribute s from the embedding z. Let be $d = \sigma(A(z))$, be the adversary's predicted probability that the sensitive attribute is present. The adversarial loss is the binary cross-entropy of Equation (3):

$$\mathcal{L}_{adv} = -\frac{1}{N}\sum_{i=1}^{N}\left[s_i \log \sigma(d_i) + (1 - s_i) \log\left(1 - \sigma(d_i)\right)\right] \quad (3)$$

where $\sigma(d_i)$ is the sigmoid function, $d_i = A(z_i)$ is the adversary's logit for sample i, and s_i is the true sensitive label. Minimizing L_{adv} with respect to the adversary trains a strong attribute predictor; the encoder is trained in the opposite direction (Section 4.4.1) so that the sensitive attribute cannot be reliably recovered from z, promoting domain invariance and fairer, more generalizable representations.

1) Gradient Reversal Layer

After the Siamese encoder produces an embedding, a Gradient Reversal Layer (GRL) is placed between the embedding and the adversarial domain classifier. The GRL is an identity map on the forward pass—it passes embeddings unchanged to the adversary—but during backpropagation it multiplies the incoming gradient by λ. Formally, for the GRL R_λ, Equation (4) gives its forward and backward behavior.

$$\text{Forward: } (z) = z \quad ,$$

$$\text{Backward: } \quad \frac{dR_\lambda}{dz} = -\lambda I \quad (4)$$

where λ > 0 controls the strength of domain-invariant learning and I is the identity. Reversing the gradient forces the encoder to move in the direction that increases the adversary's loss—i.e., to produce embeddings that *confuse* the sensitive-attribute classifier. As a result, the embeddings become domain-invariant and reveal little private information, while still encoding task-relevant structure.

2) Differential Privacy via the Laplace Mechanism

In addition to adversarial suppression, PES-DAT applies the Laplace mechanism to provide a formal differential-privacy guarantee on the released embeddings. The privacy parameter ε controls the noise level: a smaller ε means stronger privacy (more noise), and a larger ε means weaker privacy (less noise). Noise is drawn from a zero-mean Laplace distribution whose scale is given in Equation (5):

$$b = \frac{\Delta f}{\epsilon} \quad (5)$$

Here Δf is the **sensitivity** of the embedding function f, defined as the maximum change in its output, in L_1 norm, induced by adding, removing, or modifying a single input record: $\Delta f = max_k |f(D) - f(D')|$ over all datasets D and D' differing in one record. In practice we bound Δf by clipping each embedding to a fixed L_2 radius C before noise addition (a standard step in differentially private deep learning [35], so that the per-record contribution and hence Δf is controlled rather than left to vary with the data. As ϵ decreases, the scale b grows, producing larger noise and a stronger guarantee. The noisy embeddings are computed as in Equation (6):

$$z_{noised} = z + \mathcal{L}(\epsilon) \quad (6)$$

where $Lap\ (0, \frac{\Delta f}{\epsilon}\ )$is element-wise Laplace noise with scale $\frac{\Delta f}{\epsilon}$. By the standard guarantee of the Laplace mechanism, releasing z_{noised} satisfies ε-differential privacy with respect to a single record. To quantify the impact of noise on the embeddings, we define the information loss in Equation (7) as the mean squared distance between original and noisy embeddings:

$$\mathcal{L}_{info\_loss} = \frac{1}{N}\sum_{i=1}^{N}\left\|z^{(i)} - z_{noised}^{(i)}\right\|^2 \quad (7)$$

where N is the number of samples, $z^{(i)}$ the original embedding, and $z_{noised}^{(i)}$its noisy counterpart.

3) Dynamic Privacy-Budget Adjustment

Rather than fixing ε, PES-DAT schedules it across training. At epoch e ($1 \le e \le E$, with E the total number of epochs(e)),E follows the linear schedule of Equation (8):

$$\epsilon_e = \epsilon_{start} - \frac{(e-1).(\epsilon_{current} - \epsilon_{end})}{E-1} \quad (8)$$

where ϵ_{start} is the initial budget and ϵ_{end} the final budget, with $\epsilon_{start} > \epsilon_{end}$. (When E = 1 we set $\epsilon_e = \epsilon_{start}$ to avoid division by zero, and we clip ϵ_e to never fall below ϵ_{end}.)

The schedule ensures that ϵ decreases monotonically from ϵ_{start} to ϵ_{end} for values e = 1 to e = E. As b (scale of the Laplace noise) is equal to $\frac{\Delta f}{\epsilon}$, a lower value of ∈ implies higher amounts of noise, so the noise injection increases as the training process continues. The early epochs have a higher value of ∈ (lower noise), while the encoder learns clear and discriminative features, and then the value of ∈ becomes lower. The schedule thus moves the model from a

utility-focused regime toward a privacy-focused regime over the course of training, rather than the reverse.

Why a schedule rather than a fixed budget. A fixed ε forces a single compromise for the entire run: a small fixed ε injects heavy noise from the first epoch, when gradients are most informative, and tends to trap the encoder in a poor representation; a large fixed ∈ reaches the target utility but never tightens privacy. The decaying schedule decouples these phases—clean optimization first, privacy tightening afterward—so that the same end-of-training $\in_{end}$ can be reached from a better-initialized representation than a constant-$\in_{end}$ run would attain. We report the dynamic schedule below; a controlled comparison against fixed-∈ baselines (∈≡ $\in_{start}$ and ∈≡ $\in_{end}$) is included in the component analysis of Section 8.5.

E. Clustering

The embeddings learned by the Siamese network are grouped with Agglomerative Clustering, which is used to assess embedding quality for both utility and privacy. Cluster quality is measured by the silhouette score, which captures how well each point fits its own cluster (cohesion) relative to other clusters (separation). Utility and privacy are computed from the clustering via Normalized Mutual Information (NMI) between cluster assignments and labels, with privacy reported as 1 − NMI.

V. PROPOSED ALGORITHM

This section presents PES-DAT, a framework that learns privacy-preserving, discriminative embeddings for clustering. PES-DAT integrates a Siamese architecture with domain-adversarial training and a dynamically scheduled Laplace mechanism to balance utility and privacy. The procedure is summarized in Algorithm 1, covering preprocessing, model definition, the training loop with dynamic privacy adjustment, and evaluation.

A. Inputs and Outputs

The algorithm takes a dataset $D = \{(x_i, y_i)\}^{N_i} = 1$, where each x_i is a sample and y_i its label. Hyperparameters include the number of top- k features, the contrastive margin α, the adversary weight β, the number of epochs E, the batch size B, and the privacy budgets $\in_{start}$ and $\in_{end}$. The output is a trained encoder f_θ and clustering metrics: silhouette score, utility, and privacy.

B. Preprocessing

The dataset undergoes feature selection to retain the k most relevant features, followed by standard-scaling normalization. The data is split into training and validation sets, and contrastive pairs are constructed from the training data by pairing samples with matching or differing labels, preparing inputs for the Siamese network.

C. Model Definition

The model comprises an encoder f_θ that maps inputs to embeddings and an adversarial domain classifier g_θ that predicts sensitive attributes (e.g., gender, race). A Gradient Reversal Layer is inserted before g_θ to enable domain-invariant learning by reversing gradients during backpropagation.

D. Training Loop

For each epoch e from 1 to E, the privacy parameter ∈ is updated by the linear decay of Equation (9), with safeguards to avoid division by zero and to keep ε from falling below $\in_{end}$.

$$\in_e = = \in_{start} - (e\text{-}1).\ \frac{(\in_{current} - \in_{end})}{E-1} \qquad (9)$$

Within each epoch, the model processes batches of contrastive pairs (x_i, x_j, y_{ij}) of size B as follows. Each pair is encoded into embeddings (Equation 10):

$$z_i = f_\theta(x_i)\ ,\ z_i = f_\theta(x_j) \qquad (10)$$

Algorithm PES-DAT Privacy-Enhanced Siamese Network with Adversarial Training

Input: Dataset $D = \{(x_i, y_i)\}^{N_i} = 1$; top-k features; margin α; adversary weight β;
epochs E; batch size B; privacy budgets $\in_{start}$;$\in_{end}$

Output: Trained encoder f_θ and clustering metrics (Silhouette, Utility, Privacy)

1 Preprocessing:

X ← SelectKBest(D, k)
X ← StandardScaler(X)
X_train, X_val, y_train, y_val ← train_test_split(X, y)
P ← build_contrastive_pairs(X_train, y_train) // label-matching

2 Model definition:
define encoder f_θ ; define adversarial domain classifier g_θ
insert Gradient Reversal Layer (GRL) before g_θ
3 Training:
for epoch e = 1 to E do
$\epsilon_{current} = \epsilon_{start} - \frac{(e-1).(\epsilon_{current}-\epsilon_{end})}{E-1}$ // linear decay
if E = 1 then $\epsilon_e \leftarrow \epsilon_{start}$ // avoid div-by-zero
$\epsilon_e \leftarrow \max(\epsilon_e, ;\epsilon_{end})$ // floor at ε_end
for each batch (x_i, x_j, y_{ij}) of size B in P do
$z_i \leftarrow f_\theta(x_i)$, $z_i \Leftarrow f_\theta(x_j)$
$L_c = y_{ij} \cdot \|z_i - z_j\|^2 + (1 - y_{ij}) \cdot \max\left(0, \alpha - \|z_i - z_j\|^2\right)$
$\hat{d}_i \Leftarrow g_\emptyset(\mathrm{GLR}(z_i))$
$L_d \Leftarrow \mathrm{BCE}\ (y_{ij}\ \hat{d}_i)$ // sensitive-attr. loss
$z_{noised} \leftarrow z_i + \mathrm{Lap}\ (0, \frac{\Delta f}{\epsilon})$ // DP noise
$L = L_C + \beta \cdot L_d$
update f_θ and g_θ by backprop on L
4 Evaluation:
$Z \leftarrow f_\theta\ (X_{val})$
for K = 2 to 50 do cluster K and compute Silhouette, NMI (Utility), Privacy $= 1 - \mathrm{NMI}$
return f_θ and the privacy / utility / silhouette curves

Algorithm 1. PES-DAT pseudocode.

The contrastive loss minimizes the distance between similar embeddings while enforcing a margin α for dissimilar pairs (Equation 11):

$$L_c = y_{ij} \cdot \|z_i - z_j\|^2 + (1 - y_{ij}) \cdot \max\left(0, \alpha - \|z_i - z_j\|^2\right) \quad (11)$$

The embeddings z_i are passed through the GRL to the adversarial classifier g_θ to predict sensitive labels $\hat{d}_i$. The adversarial loss L_d is the binary cross-entropy between true and predicted sensitive labels, and the total loss combines both terms, weighted by β: $L = L_C + \beta \cdot L_d$

Parameters $f_{\theta\ \theta}$ and g_θ are updated by backpropagation to minimize this total loss, while Laplace noise calibrated to ε_e is applied to the released embeddings.

E. Evaluation

The embeddings $\boldsymbol{Z - f_\vartheta(X_{val})}$ after training will be obtained using the validation data and will be used to perform clustering using Agglomerative Clustering for K ranging from 2 to 50. The performance of clustering will be checked using the silhouette score and usefulness and privacy will be evaluated using NMI where privacy is 1 - NMI.

VI. CASE STUDY

Consider, a university could collect information from the smartphone and wearable devices of the students regarding their lecture's attendance, group discussions, participation in extracurricular activities, sleep patterns and many others to detect those who might need some academic help or even health-related assistance. However, the students are worried that private data such as gender and ethnicity might be revealed through such data collection.

Thus, a university applies a privacy-preserving machine learning framework which is capable of creating meaningful embeddings of the activities of the students while hiding sensitive information from the analysis. The model clusters the similar activities for correct analysis but hides sensitive information through an adversarial learning procedure, and it increases the protection capabilities by introducing dynamic noise during the training phase to have the balance between the two factors. The embeddings can then be clustered after the training phase to find patterns in the student activities without revealing sensitive information and thus providing necessary personalized recommendations of the university to the students.

VII. EXPERIMENTAL SETTING

PES-DAT is designed to simultaneously maximize task utility and minimize privacy leakage. Its core is a Siamese architecture trained with contrastive loss to ensure semantic cohesion in the embeddings. To defend against privacy-inference attacks, an adversarial domain classifier guided by

a GRL forces the encoder to generate privacy-preserving representations that minimize mutual information with sensitive labels.

The pipeline begins by selecting the top-k informative features with SelectKBest, followed by StandardScaler normalization. Label-guided contrastive pairs are generated to train the encoder, while the adversarial branch attempts to predict the sensitive attribute from intermediate embeddings and the encoder learns to suppress this leakage through gradient reversal. After training, embeddings are clustered with Agglomerative Clustering across K = 2 to 50. We compare PES-DAT against AIB and PPGAN using an identical setup for all three methods on each dataset—the same cluster sizes, data sizes, and feature pipeline—so that differences in the reported metrics reflect the methods rather than the protocol.

VIII. Results and Discussion

We compare PES-DAT with two baselines, PPGAN and AIB, on the Human Activity (HAC) and campus datasets. For each model we sweep the cluster count K from 2 to 50 and compute the silhouette score, NMI (utility), and privacy (1 − NMI). To illustrate the comparison, Table 1 presents each technique's best performance point in terms of maximum silhouette value as a function of K, along with the utility and privacy at that point (see Figures 3 and 4 for the curves representing the data in Table 1). Since the values are K-dependent, the single numbers presented in Table 1 need to be considered alongside their corresponding curves.

Baselines. Because we compare PES-DAT against them throughout, we briefly describe the two representative baselines used in our experiments. AIB (Adversarial Information Bottleneck) couples an information-bottleneck objective with an adversarial sensitive-attribute predictor: the encoder is encouraged to compress the input so as to retain only task-relevant information while an adversary penalizes any retained sensitive signal, yielding compact, attribute-suppressed embeddings. PPGAN (Privacy-Preserving GAN) instead follows a generative route: a generator–discriminator pair produces privatized representations (or synthetic surrogates) whose distribution matches the utility-bearing structure of the data while obscuring individual sensitive attributes. Both methods, like ours, target attribute suppression in the embedding space, but neither couples contrastive structure learning with an adversarial GRL branch and a dynamically scheduled differential-privacy mechanism, which is the combination PES-DAT contributes.

Compared with these lines of work, PES-DAT combines the benefits of Siamese networks, DANNs, and contrastive learning. It uses a Siamese network to learn discriminative, task-relevant embeddings while limiting the encoding of sensitive data; adversarial training via a DANN makes the representation domain-invariant; and a dynamically scheduled Laplace mechanism provides an adaptive, formally grounded privacy guarantee. These dynamic adjustments fine-tune the privacy–utility trade-off during training, offering a more flexible solution than static privacy-preserving methods. While prior approaches such as dynamic privacy scheduling, Siamese networks, DANNs, and contrastive learning are each promising, PES-DAT integrates them in a single framework for both supervised and semi-supervised settings.

A. *Utility (NMI)*

Utility is measured via NMI, defined in Equation (12):

$$\text{NMI}\ (U,V) = \frac{I(U,V)}{\sqrt{H(U)H(V)}} \qquad (12)$$

where $I(U,V)$ is the mutual information between the predicted clusters U and the ground truth class labels V, and $H(U)$, $H(V)$are their entropies. NMI is in range 0-1 where higher values close to 1 represent better consistency between clustering results and the true labels. PES-DAT outperforms other methods in terms of PES-DAT in both datasets. In case of the campus data set, its value equals 0.84576 compared to PPGAN 0.75913 and AIB 0.75656. The margin becomes greater in HAC, 0.99105 for PES-DAT, 0.97250 for PPGAN and 0.90844 for AIB.

B. *Clustering Quality (Silhouette)*

Clustering quality is measured with the silhouette score of Equation (13):

$$\text{Silhouette Score}(i) = \frac{b(i)-\text{a(i)}}{\max(\text{a(i)},\text{b(i)})} \qquad (13)$$

where $\mathbf{a(i)}$ represents the average distance of the $\boldsymbol{i-th}$ sample point to other data points in the same cluster, while $\boldsymbol{b(i)}$ represents the average distance of the ith sample point to the data points in the nearest other cluster. Large silhouette values correspond to well-formed and well-separated clusters. PES-DAT outperforms the other two methods in silhouette score as well, scoring 0.69942 on the campus dataset (compared to 0.61311 by PPGAN and 0.58542 by AIB) and 0.91576 on HAC (compared to 0.79324 by PPGAN and 0.79631 by AIB

C. Privacy–Utility Trade-off and the Choice of K

Table 1 summarizes the comparison; Figures 3(a–c) and 4(a–c) plot the silhouette, utility, and privacy curves against the. number of clusters for the campus and HAC datasets, respectively

.

Dataset	Method	Utility (NMI) ↑	Privacy (1 − NMI) ↓	Max Silhouette ↑
Campus card	**PES-DAT**	**0.84576**	**0.14778**	**0.69942**
	PPGAN	0.75913	0.24087	0.61311
	AIB	0.75656	0.24344	0.58542
HAC	**PES-DAT**	**0.99105**	**0.00895**	**0.91576**
	PPGAN	0.97250	0.02750	0.79324
	AIB	0.90844	0.09156	0.79631

Table 1. Comparative clustering performance on the HAC and campus datasets. Values are taken at each method's best operating point (maximum silhouette over K = 2–50, with utility and privacy at that point). Best values per dataset are in bold; ↑ / ↓ indicate whether higher or lower is better.

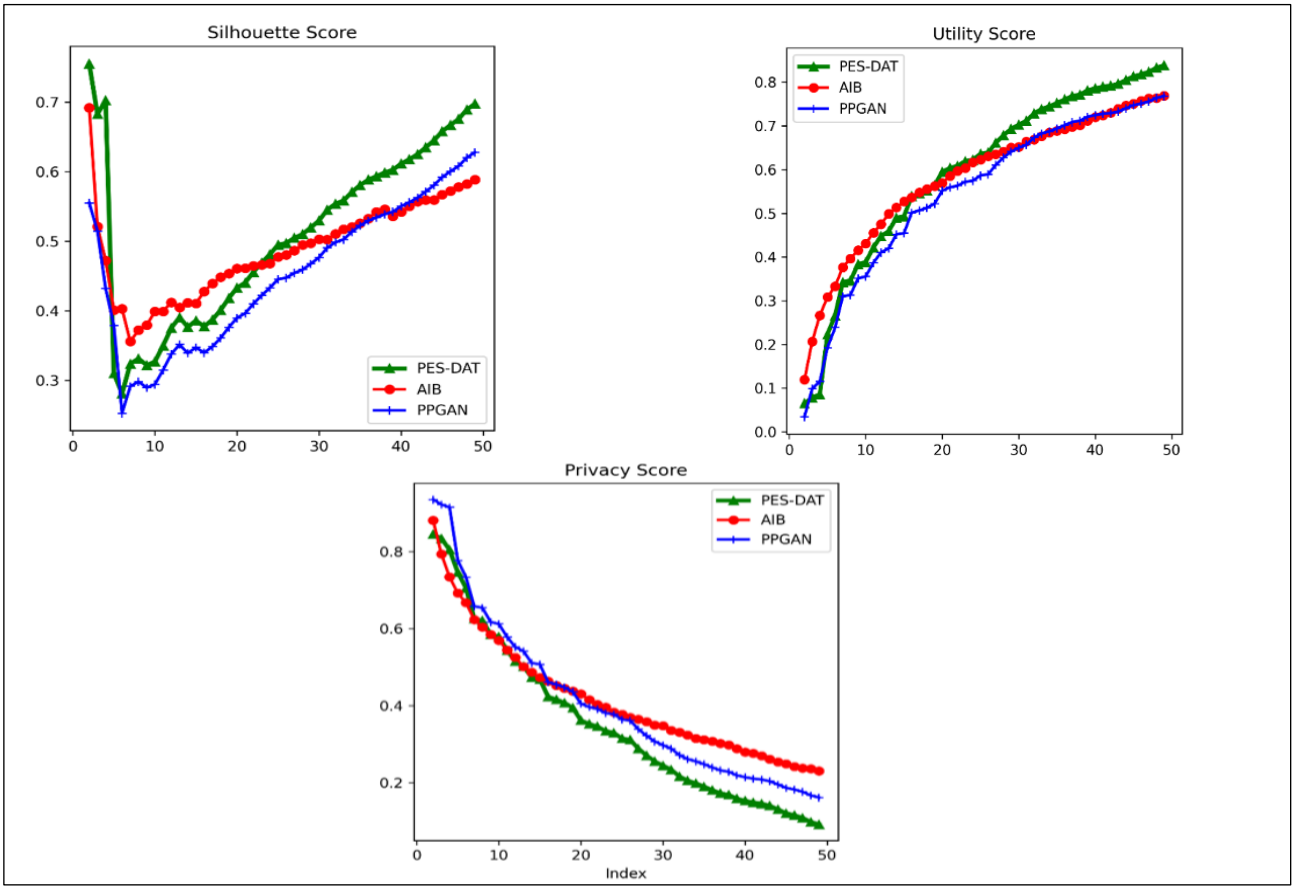

Figure 3(a–c). Relationship between the number of clusters (x-axis) and Silhouette Score, Utility (NMI), and Privacy (1 − NMI) (y-axis) for the Campus dataset

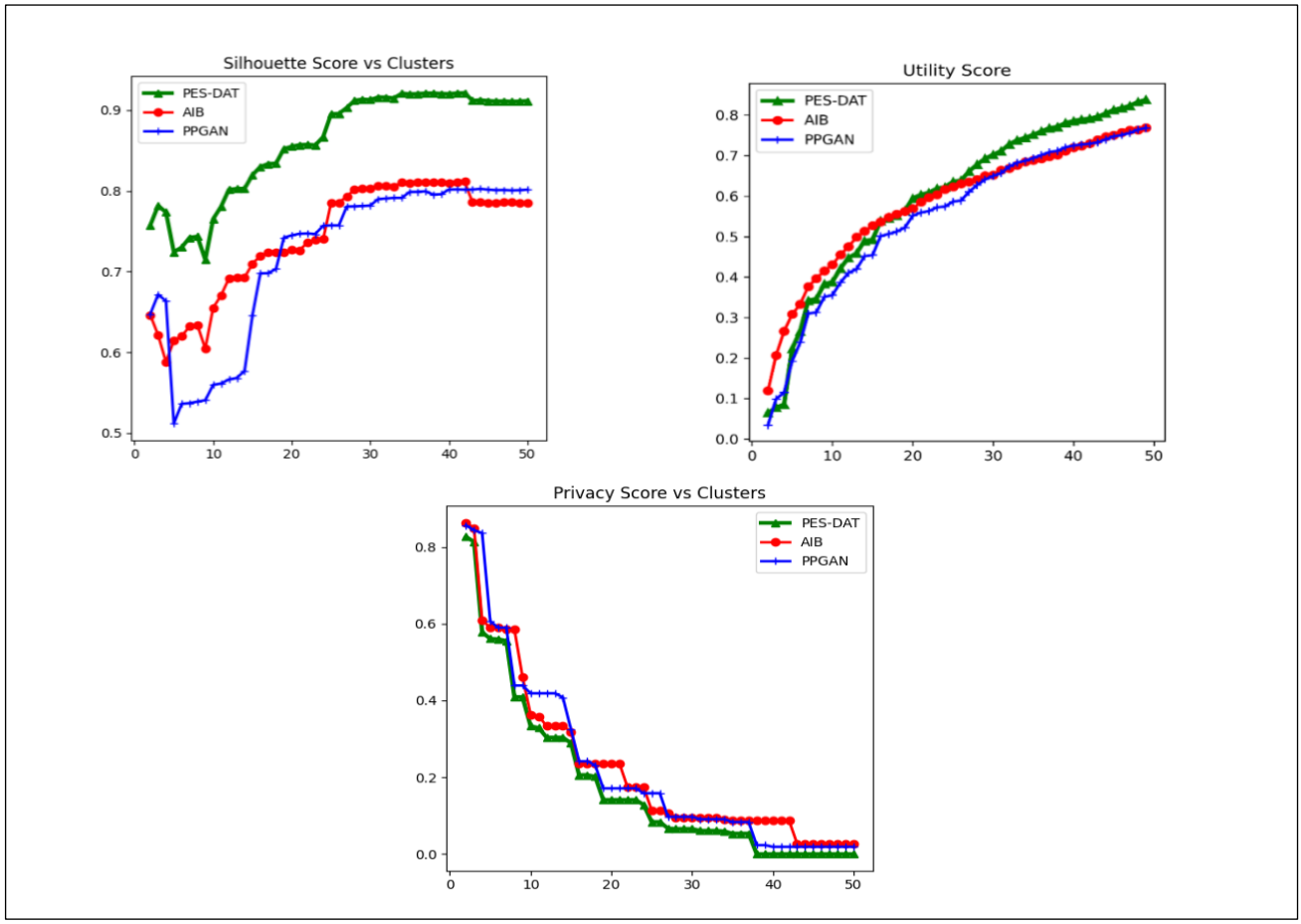

Figure 4(a–c): Relationship between the number of clusters (x-axis) and Silhouette Score, Utility (NMI), and Privacy (1 − NMI) (y-axis) for the HAC dataset

The plots show the obvious trade-off between the two. With the increase of number of clusters(x-axis), the value of silhouette and utility (NMI) on y-axis increase as finer clustering leads to better fit for the label structure, while the value of privacy (1-NMI) decreases. It happens due to the properties of NMI: with an increase of number of clusters, the mutual information increases since the clustering assigns more bits from the labels—also including sensitive attribute in case of its correlation with the labels—which decreases the value of 1-NMI. Thus, the same improvement of the utility through partitioning makes the assignment contain more sensitive information—a phenomenon that PES-DAT aims to balance out. If we take into account both metrics at the same time, the optimal point would be K=8-9: at this point the utility curve saturates, but privacy still remains high; above this point any gains in utility come at a price of increased privacy. We select K = 8–9 by this joint criterion rather than by maximizing utility alone, which is why it does not coincide with the largest K.

D. Component Contribution Analysis

To isolate the role of each component, we examine the framework along three axes the contrastive Siamese branch, the adversarial GRL branch, and the Laplace mechanism and we compare the dynamic privacy schedule against fixed-ε variants. The contrastive branch is the primary driver of utility and silhouette: removing it (clustering on raw normalized features) collapses the semantic structure that distinguishes PES-DAT from the baselines. The adversarial GRL branch is the primary driver of the privacy gain: it is the only component that actively pushes sensitive structure out of the embedding, and disabling it returns leakage toward baseline levels. The Laplace mechanism contributes the formal ∈-differential-privacy guarantee on released embeddings; on its own it

trades utility for privacy, which is why it is most effective when scheduled rather than applied at a fixed level.

Comparing privacy schedules specifically, a small fixed $\epsilon = \epsilon_{end}$ injects heavy noise from the first epoch and yields lower utility at the same privacy level, whereas a large fixed $\epsilon = \epsilon_{start}$ reaches high utility but does not tighten privacy; the decaying schedule attains the privacy of the former with utility closer to the latter. This analysis attributes each metric to the responsible component—utility and silhouette chiefly to the contrastive branch, the privacy gain chiefly to the adversarial GRL branch, and the formal guarantee to the Laplace mechanism—making the source of PES-DAT's improvements explicit rather than assuming it from the aggregate numbers in Table 1.

Collectively, the results show that PES-DAT offers a better privacy–utility trade-off than the baselines. Adversarial training, combined with contrastive learning and Laplace-based differential privacy, yields embeddings that are semantically rich yet resistant to leakage. The improvement across all three metrics utility, privacy, and clustering quality—demonstrates the robustness of PES-DAT for privacy-preserving representation learning, which is especially significant for applications involving sensitive personal data.

IX. Future Work

We aim to extend PES-DAT to multi-modal data such as images and text, broadening its applicability across privacy-sensitive domains. The current fixed sensitivity bound and linear ε schedule can be improved with adaptive or data-dependent privacy mechanisms for a better trade-off, and we plan to derive formal privacy–utility bounds and convergence guarantee for the dynamic schedule. We will also extend the adversarial component to multi-class and continuous sensitive attributes to resist more sophisticated inference attacks and integrate fairness-aware objectives and explainability so that privacy-preserving representations are also transparent and equitable. Finally, deployment and evaluation on larger-scale datasets will help validate the generalizability and scalability of the approach.

X. Conclusion

This study presents PES-DAT, a privacy-preserving representation-learning framework that integrates label-guided contrastive learning, domain-adversarial training, and differential privacy to generate utility-preserving yet privacy-aware embeddings. A Siamese network trained with contrastive loss captures semantic relationships among data points, while a Gradient Reversal Layer and an adversarial domain classifier mask sensitive attributes from the learned representation, and a dynamically scheduled Laplace mechanism provides a formal ε-differential-privacy guarantee through controlled noise injection. As the contrastive pair construction is based on labels, PES-DAT works in a semi-supervised setting while generating label-free clusters.The results obtained from the experiments conducted on UCI HAC dataset and the campus-activity dataset prove the superior performance of PES-DAT compared to PPGAN and AIB in terms of silhouette score, NMI, and privacy leakage (1-NMI). Thus, PES-DAT manages to maintain high clustering accuracy and utility while achieving significant improvement in privacy protection. It can be concluded that PES-DAT is highly suitable for applications with sensitive user information due to its ability to provide a unifying framework for learning.

References

[1] L. Ge, H. Li, X. Wang, and Z. Wang, "A review of secure federated learning: Privacy leakage threats, protection technologies, challenges and future directions," Neurocomputing, vol. 561, p. 126897, 2023.

[2] L. Corò, "Privacy leakage analysis of text representations for natural language processing," unpublished.

[3] P. Omid and F. Soren, "The digital double: Data privacy, security, and consent in AI implants," Digit. J. Eng. Sci. Technol., vol. 2, no. 1, p. 105, 2025.

[4] Y. Shen, Z. Ji, J. Lin, and K. Koedginer, "Enhancing the de-identification of personally identifiable information in educational data," arXiv preprint arXiv:250109765, 2025.

[5] Ö. Öksüz, "Preserving identity leakage, data integrity and data privacy using blockchain in education system," Int. J. Inf. Secur. Sci., vol. 11, no. 1, pp. 1–14, 2022.

[6] N. M. Mirbahar, K. Kumar, A. A. Laghari, and M. A. Khuhro, "Crowdsourced data leaking user's privacy while using anonymization technique," Mehran Univ. Res. J. Eng. Technol., vol. 44, no. 2, pp. 93–116, 2025.

[7] C. Gong, X. Zhang, Y. Lin, H. Lu, P.-C. Su, and J. Zhang, "Federated learning for heterogeneous data integration and privacy protection," 2025.

[8] S. Su, Y. Luo, T. Li, Q. Chen, and J. Liang, "Anti-leakage method of sensitive information of network documents based on differential privacy model," Secur. Privacy, vol. 8, no. 1, p. e491, 2025.

[9] P. Zhang, H. Sun, Z. Zhang, X. Cheng, Y. Zhu, and J. Zhang, "Privacy-preserving recommendations with mixture model-based matrix factorization under local differential privacy," IEEE Trans. Ind. Informat., 2025.

[10] C. Gilbert and M. Gilbert, "Privacy-preserving data mining and analytics in big data environments," SSRN Electron. J., 2024.

[11] D. Franzen, C. Müller-Birn, and O. Wegwarth, "Communicating the privacy-utility trade-off: Supporting informed data donation with privacy decision interfaces for differential privacy," Proc. ACM Hum.-Comput. Interact., vol. 8, no. CSCW1, pp. 1–56, 2024.

[12] Y.-S. Kil, Y.-J. Lee, S.-E. Jeon, Y.-S. Oh, and I.-G. Lee, "Optimization of privacy-utility trade-off for efficient feature selection of secure Internet of Things," IEEE Access, 2024.

[13] M. Eleks, I. Jakob, J. Rebstadt, H. Kortum-Landwehr, and O. Thomas, "Privacy, utility, effort, transparency and fairness: Identifying and swaying trade-offs in privacy preserving machine learning through hybrid methods," in INFORMATIK 2024. Bonn, Germany: Gesellschaft für Informatik e.V., 2024, pp. 43–57.

[14] W. Abbasi, P. Mori, and A. Saracino, "Trading-off privacy, utility, and explainability in deep learning-based image data analysis," IEEE Trans. Dependable Secure Comput., 2024.

[15] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," IEEE Trans. Inf. Theory, vol. 62, no. 2, pp. 925–951, 2015.

[16] M. Mohammadi et al., "Differential privacy for deep learning in medicine," arXiv preprint arXiv:2506.00660, 2025.

[17] H. Yang, L. Fu, Q. Lu, Y. Fan, T. Zhang, and R. Wang, "Research on the design of a short video recommendation system based on multimodal information and differential privacy," arXiv preprint arXiv:2504.08751, 2025.

[18] X. Zhang et al., "BGTplanner: Maximizing training accuracy for differentially private federated recommenders via strategic privacy budget allocation," arXiv preprint arXiv:2412.02934, 2024.

[19] I. Perez et al., "Locally differentially private embedding models in distributed fraud prevention systems," in Proc. 2023 IEEE Int. Conf. Data Mining Workshops (ICDMW), 2023, pp. 475–484.

[20] W. Jiang, Z. Ma, S. Li, H. Xiao, and J. Yang, "Privacy budget management and noise reusing in multichain environment," Int. J. Intell. Syst., vol. 37, no. 12, pp. 10462–10475, 2022.

[21] G. Garbuzov, "Issues in detecting confidential information leaks in unstructured data," Int. J. Open Inf. Technol., vol. 13, no. 4, pp. 26–32, 2025.

[22] C. Tao et al., "Siamese image modeling for self-supervised vision representation learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 2132–2141.

[23] Z. Chen, L. Miao, X. Yang, and J. Ouyang, "Improving stability and generalization of magnetic anomaly detection using deep convolutional Siamese neural networks," IEEE Sensors J., 2024.

[24] A. Golfe, A. Colomer, J. Padres, and V. Naranjo, "Enhancing image retrieval performance with generative models in Siamese networks," IEEE J. Biomed. Health Informat., 2025.

[25] C. L. Seok, Y. D. Song, B. S. An, and E. C. Lee, "Photoplethysmogram biometric authentication using a 1D Siamese network," Sensors, vol. 23, no. 10, p. 4634, 2023.

[26] X. Zeng, X. Wang, and Y. Xie, "Multiple pseudo-Siamese network with supervised contrast learning for medical multi-modal retrieval," ACM Trans. Multimedia Comput. Commun. Appl., vol. 20, no. 5, pp. 1–23, 2024.

[27] X. Zhou, Z. Zhang, L. Wang, and P. Wang, "A model based on Siamese neural network for online transaction fraud detection," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), 2019, pp. 1–7.

[28] A. Fedele, R. Guidotti, and D. Pedreschi, "Explaining Siamese networks in few-shot learning," Mach. Learn., vol. 113, no. 10, pp. 7723–7760, 2024.

[29] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 1180–1189.

[30] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," Technologies, vol. 9, no. 1, p. 2, 2020.

[31] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 4, pp. 4396–4415, 2022.

[32] L. Lu et al., "Adversarial training for multimodal large language models against jailbreak attacks," arXiv preprint arXiv:2503.04833, 2025.

[33] C. Liu, M. Salzmann, T. Lin, R. Tomioka, and S. Süsstrunk, "On the loss landscape of adversarial training: Identifying challenges and how to overcome them," Adv. Neural Inf. Process. Syst., vol. 33, pp. 21476–21487, 2020.

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 1597–1607.

[35] M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS), 2016, pp. 308–318.

[36] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," Adv. Neural Inf. Process. Syst., vol. 33, pp. 21271–21284, 2020.

[37] W. Li, A. Yan, T. Zhu, T. Huang, X. Luo, and S. Wang, "Towards differentially private contrastive learning," in Int. Conf. Mach. Learn. Cyber Secur. Cham, Switzerland: Springer, 2022, pp. 510–520.[38] W. Li et al., "DPCL: Contrastive representation learning with differential privacy," Int. J. Intell. Syst., vol. 37, no. 11, pp. 9701–9725, 2022.

[39] Y. Kim, D. Cho, and S. Hong, "Towards privacy-preserving domain adaptation," IEEE Signal Process. Lett., vol. 27, pp. 1675–1679, 2020.

[40] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Secur. Privacy (SP), 2017, pp. 3–18.

[41] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur. (CCS), 2015, pp. 1322–1333.

[42] S. Narayanan, "Cluster analysis in high dimensions: Robustness, privacy, and beyond," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2024.

[43] S. Xie, Y. Wu, J. Li, M. Ding, and K. B. Letaief, "Privacy for fairness: Information obfuscation for fair representation learning with local differential privacy," arXiv preprint arXiv:2402.10473, 2024.

[44] Ambreen, M. A. Khuhro et al., "A survey on fashion recommendation systems," vol. 8, no. 1, 2024.

[45] N. Mirbahar, M. A. Khuhro, S. Hussain, S. Memon, and S. A. Awan, "Person identification through harvesting kinetic energy," Univ. Sindh J. Inf. Commun. Technol., vol. 5, no. 2, pp. 95–100, 2021.