RESEARCH PAPER

Learning Task-Preserving Encoded Representations for Privacy-Sensitive Data Outsourcing

RADHA V [1*], VARSHINI M [2*], ZAHIRA SHIRIN S [3*], YOGA PRAPANIJANI P [4*], SUBASRI S [5*]

1 Assistant Professor, Department of Computer Science and Engineering, V.S.B College of Engineering Technical campus, kinathukadavu, Coimbatore, Tamilnadu, India.*

2 UG Scholar , Department of Computer Science and Engineering, V.S.B College of Engineering Technical campus, kinathukadavu, Coimbatore, Tamilnadu, India.*

3 UG Scholar, Department of Computer Science and Engineering, V.S.B College of Engineering Technical campus, kinathukadavu, Coimbatore, Tamilnadu, India.*

4 UG Scholar, Department of Computer Science and Engineering, V.S.B College of Engineering Technical campus, kinathukadavu, Coimbatore, Tamilnadu, India.*

5 UG Scholar, Department of Computer Science and Engineering, V.S.B College of Engineering Technical campus, kinathukadavu, Coimbatore, Tamilnadu, India.*

***Abstract*—**This paper addresses a fundamental challenge in modern machine learning: how to train accurate cloud-hosted models on sensitive data without ever exposing that data to the cloud. It introduces LTPER (Learning Task-Preserving Encoded Representations), a two-zone framework in which a Supervised Residual Autoencoder (SRAE) is trained entirely on the data owner's local hardware and only the resulting encoded artifact Ψ is forwarded to the cloud for downstream model training. The encoder is optimized simultaneously under four objectives designed to work in concert: a reconstruction term that anchors the encoding to input geometry, a multi-layer classification term that enforces task relevance at every encoder depth, a center loss that tightens intra-class clusters, and a PCA cosine alignment term that aligns the encoding topology with the principal variance directions of the original data. This paper further introduces a layer-stacking strategy that concatenates encoder outputs from all depth levels into a single 480-dimensional vector, recovering multi-scale structure that single-bottleneck approaches discard. LTPER is validated on five benchmark datasets spanning images, medical scans, gene-expression arrays, and multi-omics records. The encodings match or exceed unprotected baselines across four classifier families while resisting reconstruction attacks. This paper also provides a four-scenario adversary analysis that formally establishes the security boundaries of the proposed system.

***Keywords*—**privacy-preserving machine learning; supervised residual autoencoder; task-preserving representation; Supervised Residual Autoencoder; PCA cosine alignment; secure data outsourcing; multi-objective optimization; Learning Task-Preserving Encoded Representations

How to cite this article: Radha V, Varshini M, Zahira Shirin S, Yoga Prapanijani P, Subasri S. Learning task-preserving encoded representations for privacy-sensitive data outsourcing. Int J Drug Deliv Technol. 2026;16(74s): 300-307. DOI: 10.25258/ijddt.16.74s.34

.

I. INTRODUCTION

Organizations across healthcare, genomics, finance, and manufacturing increasingly depend on cloud platforms to train machine learning models at scale. On-premise GPU infrastructure is costly and scarce, pushing institutions toward third-party services for computationally demanding workloads. Cloud-based machine learning has also demonstrated practical value in healthcare contexts, such as AI-driven diabetic risk prediction systems that leverage ensemble learning over cloud infrastructure [1]. At the same time, the datasets that feed these workloads are among the most sensitive imaginable: patient records, genomic sequences, financial transactions, and biometric logs. Regulations such as GDPR, HIPAA, and CCPA impose strict constraints on where such data can travel and who can access it. Simply uploading raw records to a cloud training pipeline is not an option: doing so opens the door to model inversion attacks that can reconstruct training samples from model weights, and to membership inference attacks that can determine whether a specific individual's data was used during training [2], [3], [4].

The research community has pursued three main lines of defense. Homomorphic encryption (HE) allows arithmetic to be performed directly on ciphertexts, so the cloud never sees raw values, but the computational overhead is severe—often 100 to 10,000 times that of unprotected training—and non-linear activations require expensive polynomial approximations [4]. Differential privacy (DP) adds carefully calibrated noise to gradients during training, providing a formal bound on information leakage, but the accuracy cost compounds rapidly as feature dimensionality grows [3]. Beyond these cryptographic defenses, big data analytics combined with deep learning has been applied to large-scale cloud pipelines for tasks such as sentiment classification, illustrating both the power and the privacy exposure of cloud-hosted models [5].

Federated learning (FL) keeps raw data on each local device and shares only gradient updates, yet those gradients themselves can be inverted to reconstruct training inputs, and

the approach is further vulnerable to Byzantine manipulation and communication bottlenecks [6], [3]. Complementary work has shown that ML pipelines can be tuned through real-time user feedback signals [7]. Together, these efforts highlight that securing cloud ML requires coordinated protection at both the data layer and the system layer—no single mechanism is sufficient.

This paper takes a different approach. Instead of encrypting raw data or adding noise to gradients, an encoder is trained locally that converts each raw record into a compact representation designed to be useful for downstream classification but impossible to invert into the original data without access to the encoder itself. Only these representations are sent to the cloud. Prior work explored similar ideas using a simple supervised autoencoder [8], [9], [10], but those solutions shared only the bottleneck layer, discarding the richer intermediate structure captured by earlier encoder stages. They also lacked a formal security model. LTPER is designed to overcome both limitations.

The specific contributions of this paper are:
(i) This paper designs an SRAE trained under four complementary loss terms that together produce encodings that are simultaneously accurate, compact, and geometrically aligned with the source data [11], [12];
(ii) This paper introduces a layer-stacking strategy that concatenates encoder outputs across all depth levels into a fixed 480-dimensional representation, recovering hierarchical structure lost at the bottleneck [13], [10];
(iii) This paper validates LTPER on five heterogeneous benchmarks with full per-objective ablation [14], [11];
(iv) This paper provides a four-scenario adversary model that formally characterizes the security of the framework and identifies residual risks [2], [3].

II. Related Work

Privacy-preserving machine learning has become an important research area due to the increasing need to process sensitive data in distributed or cloud environments. Several studies have explored different techniques to ensure data privacy while maintaining model utility. Chiang [13] proposed a privacy-preserving convolutional neural network training method using transfer learning and hidden-layer encoding to protect sensitive training data. Similarly, Quintero Ossa et al. [8] introduced a collaborative data-sharing framework that utilizes autoencoder latent space embeddings to enable privacy-preserving machine learning without exposing raw data.

Comprehensive surveys have also been conducted to evaluate existing privacy-preserving machine learning techniques. Zhang and Li [2] presented a detailed analysis of approaches for improving privacy in machine learning datasets, highlighting methods such as differential privacy, encryption, and representation learning. Raja [9] further explored collaborative data sharing using autoencoder latent space embeddings to protect sensitive information during model training.

Adversarial learning and representation-based privacy techniques have also been studied extensively. Liguori et al. [15] proposed adversarial regularized reconstruction methods for anomaly detection that improve the robustness of generated representations while limiting data leakage. Azizian and Bajić [10] developed a privacy-preserving autoencoder framework for collaborative object detection, demonstrating that encoded feature representations can maintain task performance while preventing access to original images.

Network security and system-level protection mechanisms have also been investigated in the literature. Manivannan et al. [16] proposed adaptive hierarchical machine learning techniques for real-time DDoS detection in distributed systems, demonstrating scalable adversary defense strategies that are directly relevant to protecting cloud-hosted ML pipelines. Similarly, Mary et al. [17] investigated adversary identification in network environments using piggybacking protocols, providing insights into how adversarial behavior can be detected and attributed in distributed cloud architectures. Pathak et al. [7] applied advanced machine learning methods to analyze user feedback on virtual assistants to improve system performance and optimization in intelligent environments. Sathyanarayanan et al. [18] designed a cybersecurity framework for encrypted file management systems to ensure secure storage and controlled access to sensitive digital resources.

Cryptographic approaches also play a crucial role in privacy-preserving machine learning. Zhai et al. [19] proposed privacy-preserving outsourcing algorithms for multidimensional data encryption in smart grid systems, enabling secure processing of sensitive energy data. Bandhela et al. [14] developed an optimized deep learning-based privacy-preserving approach for secure data mining, demonstrating improved security and performance in data-driven applications. Chen et al. [4] provided a comprehensive survey of cryptography-based privacy-preserving machine learning techniques, emphasizing the role of encryption methods such as homomorphic encryption and secure multiparty computation.

Adaptive privacy management frameworks have also been explored to dynamically balance privacy and utility. Hui et al. [20] introduced PrivacyPAD, a reinforcement learning framework that dynamically manages privacy-aware delegation strategies. Complementing this, Mary et al. [21] demonstrated that hybrid neural networks can effectively detect adversarial and deceptive content in online environments, underscoring the applicability of multi-model architectures in security-critical classification tasks analogous to the multi-objective encoding pursued by LTPER. Ouaari et al. [11], [12] proposed a multi-objective autoencoder approach for robust representation learning that simultaneously optimizes data utility and privacy protection.

Federated learning has emerged as another promising paradigm for privacy-preserving machine learning. He and Joshi [6] proposed PPFL-RDSN, a federated learning-based residual dense spatial network that enables privacy-preserving image reconstruction while keeping training data distributed

across devices. Zeng et al. [3] conducted a systematic review of efficient privacy-preserving machine learning systems and highlighted federated learning, differential privacy, and cryptographic computation as the most promising research directions.

In addition, domain-specific privacy-preserving solutions have been developed for industrial environments. Terziyan et al. [22] proposed privacy-preserving image encryption and generation techniques for smart manufacturing applications, ensuring that sensitive visual data remains protected during machine learning processes.

Overall, existing research demonstrates significant progress in privacy-preserving machine learning through encryption, federated learning, differential privacy, and representation-based methods. However, many approaches still suffer from computational overhead, reduced model performance, or limited protection against advanced inference attacks. These limitations motivate the development of more efficient frameworks capable of preserving both data privacy and task performance in cloud-based machine learning environments.

II. PROPOSED METHODOLOGY

This paper designs LTPER around a strict two-zone separation, illustrated in Fig. 1. Everything that touches raw data—preprocessing, encoding, adversarial decoding, and classification supervision—runs exclusively on the data owner's on-premise hardware. The cloud receives only the encoded dataset Ψ and uses it to train downstream classifiers. No raw record, weight matrix, or gradient ever crosses the zone boundary.

A. Problem Formulation

The owner's dataset is denoted as $D = \{(x_i, y_i)\}_{i=1}^{R}$, where $x_i \in \mathbb{R}^d$ is a d-dimensional record and $y_i \in \{1,\ldots,C\}$ its label. The goal is to find an encoder $E : \mathbb{R}^d \to \mathbb{R}^m$ ($m \ll d$) and a surrogate dataset $\Psi = \{(\psi_i, y_i)\}$ that simultaneously satisfies three properties: (P1) a classifier trained on Ψ achieves accuracy on par with one trained on D directly; (P2) recovering x_i from ψ_i without $E(\cdot)$ is computationally infeasible; and (P3) ψ_i reveals nothing about the original data type, modality, or statistical distribution. Crucially, the transmitted object ψ_i is not a noisy or encrypted copy of x_i—it is an entirely new mathematical artifact produced by a learned transformation that the cloud never sees.

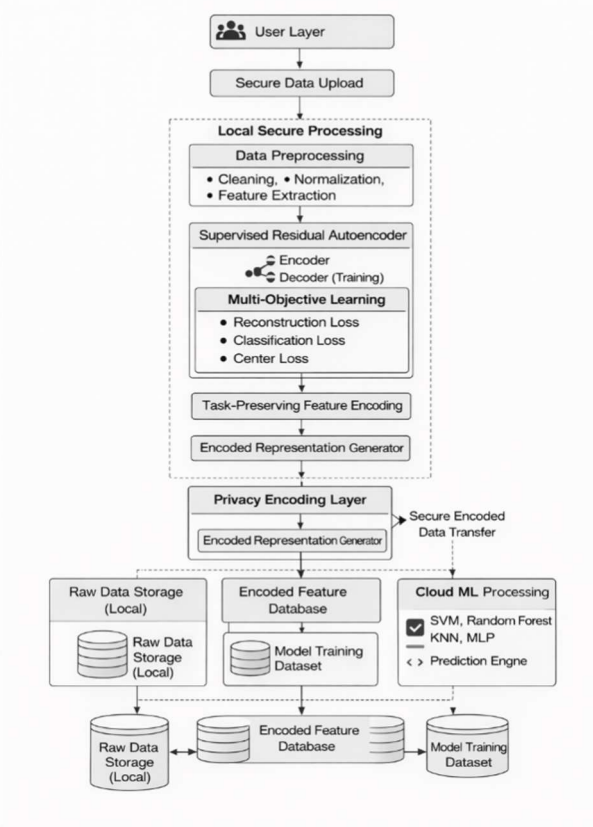

Fig. 1. LTPER two-zone architecture. Left: on-premise SRAE training. Right: cloud classifier fitting on Ψ. Raw data never crosses the privacy boundary.

B. Supervised Residual Autoencoder (SRAE)

The heart of the local pipeline is the SRAE, which consists of a depth-L residual encoder $E(\cdot)$, an adversarial decoder $D(\cdot)$ used only during local training, and L lightweight classification probes {f_j}_{j=1}^L attached to each encoder layer. Residual skip connections are used so each block learns an additive correction rather than a full mapping, which enables training deep stacks without gradient vanishing:

$$y = F(x; W) + x \quad (1)$$

For image inputs (MNIST, FashionMNIST, Retinal OCT) Convolutional blocks are instantiated with 3×3 kernels, batch normalization, and ReLU activations (C-SRAE). For tabular and genomic inputs (Leukemia, TCGA) fully-connected blocks are used with ELU activations and L2 regularization (SRAE). The decoder mirrors the encoder in transposed form and is used solely to compute reconstruction loss; it is never shared with the cloud. A MINE-based Mutual Information Monitor to each encoder layer to track and constrain how much raw-input information each stage retains, preventing early layers from accumulating reconstructible signal.

C. Layer-Stacking Encoding Strategy

A key design decision in this framework is what to transmit to the cloud. Sending only the bottleneck output loses the hierarchical structure built up by earlier encoder stages. Instead, this paper concatenates the output of every encoder layer along the feature dimension:

$$\psi_i = [\psi_i^{(1)T}, \psi_i^{(2)T}, \ldots, \psi_i^{(L)T}]^T \in \mathbb{R}^{480} \quad (2)$$

Early layers encode fine-grained local patterns such as edge statistics and gene co-expression signatures; deeper layers encode high-level semantic prototypes. Stacking all levels into a 480-dimensional vector—allocated as [160, 120, 100, 100] across fthe encoder blocks—gives downstream classifiers access to both granular and abstract cues simultaneously. A secondary benefit is improved inversion resistance: because ψ_i does not correspond to the input of any single decoder layer, an adversary cannot reconstruct x_i by inverting one bottleneck; they would have to reconstruct the entire encoder pathway.

D. Four-Objective Loss Function

The SRAE is jointly optimized under the following weighted loss:

$$L_{total} = \lambda_r L_r + \lambda_c L_{cls} + \lambda_k L_{center} + \lambda_p L_{pca} \quad (3)$$

The weights $(\lambda_r, \lambda_c, \lambda_k, \lambda_p) = (1.0, 0.8, 0.5, 0.3)$ are determined via grid search on a held-out validation fold and fixed across all experiments.

1) Reconstruction Loss L_r:

Mean squared error is measured between each input x_i and its decoder reconstruction $\hat{x}_i = D(E(x_i))$. This term prevents the encoder from collapsing all inputs to a constant vector and controls how tightly the encoding preserves input geometry:

$$L_r = (1/N) \Sigma_i \|x_i - D(E(x_i))\|_2^2 \quad (4)$$

2) Classification Loss L_cls:

Cross-entropy is summed across all L layer-wise classification probes rather than only the final layer. This forces every encoder stage to maintain task-discriminative structure, ensuring that all 480 dimensions of ψ_i carry predictive information:

$$L_{cls} = -(1/NL) \Sigma_j \Sigma_i y_i \log f_j(\psi_i^{(j)}) \quad (5)$$

3) Center Loss L_center:

Each encoding is pulled toward the running centroid of its class, compressing within-class scatter while L_cls handles between-class separation:

$$L_{center} = \Sigma_i \| \psi_i - c_{y_i} \|_2^2 \quad (6)$$

Centroids c_k are maintained as exponential moving averages, avoiding the memory cost of an explicit memory bank. This term is especially valuable for genomic datasets where class boundaries are biologically subtle and within-class variance is high.

4) PCA Cosine Alignment Loss L_pca:

A two-unit projection head f_2 is attached to the encoding and align its output with the 2D PCA projection of the original input via cosine similarity. This grounds the encoding topology in the dominant variance structure of the source domain:

$$L_{pca} = \Sigma_i [1 - \cos(f_2(\psi_i), x_i^{pca})] \quad (7)$$

For multi-omics datasets, PCA axes correspond to biologically interpretable variance components, so this term also acts as a regularizer that stabilizes training when sample counts are small.

E. Datasets and Training Configuration

This paper evaluates on five benchmarks that span three modalities and represent a broad range of privacy-sensitive settings: (1) MNIST— 70,000 grayscale handwritten digit images, 10 classes; (2) FashionMNIST— 70,000 apparel images with ten semantically overlapping categories, testing the encoder under visual ambiguity; (3) Retinal OCT— 84,495 clinical retinal scans across four diagnoses (Normal, DME, Drusen, CNV), a dataset with direct patient re-identification risk; (4) Leukemia Gene Expression— 281 patients × 22,284 gene probes across 7 subtypes, the most challenging high-dimensional case in this study; and (5) TCGA Breast Cancer Multi-Omics— three modalities (DNA mutation, RNA-seq, methylation) held by separate institutions, exercising the vertical federation encoding path of LTPER.

All SRAE variants are trained using Adam with an initial learning rate of 10^{-3} halved every 30 epochs, early stopping at patience 20 on validation reconstruction loss, and L1 kernel regularization at 10^{-5}. Batch sizes of 64 are used for images and 32 for tabular inputs. Once training converges, Ψ is transmitted to the cloud under AES-256 encryption via the Secure Feature Sharing module. Cloud classifiers (KNN, SVM, Random Forest, MLP) are then fitted on Ψ using 10-fold stratified cross-validation with randomized hyperparameter search. At no point during this process do the raw dataset, encoder weights, or decoder weights leave on-premise storage. The local training overhead is 2×–5× over a standard single-objective run; cloud inference latency is unaffected.

F. Threat Model and Security Guarantees

This paper analyzes LTPER under the honest-but-curious adversary model, in which each party follows the protocol faithfully but attempts to infer sensitive information from everything they legitimately receive. The encoder $E(\cdot)$ stays on-premise and is never transmitted under any scenario. Four adversary classes are considered:

Cloud Provider: The cloud sees Ψ and the trained model M_Ψ but not $E(\cdot)$. Without $E(\cdot)$, it cannot determine the original dimensionality, modality, or input distribution. Recovering x_i from ψ_i is an underdetermined inverse problem with no unique solution. It is recommended to disable soft-probability outputs at the serving API to reduce residual membership inference risk.

End User (API Access): API users receive only predicted labels. Without Ψ or $E(\cdot)$, gradient-based inversion is impossible. Label-only boundary attacks require thousands of adaptive queries and yield only coarse approximations of the decision boundary.

Collaborating Institution: A peer institution may hold $E(\cdot)$ and M_Ψ but is never given Ψ directly. Reconstructing a training sample requires optimizing a candidate input toward a

target encoding ψ_i, which is withheld. It is recommended to apply rate-limiting and embed query-tracing watermarks in Ψ to mitigate model extraction.

Colluding Cloud + Institution: This is the highest-risk scenario. Joint possession of Ψ and $E(\cdot)$ enables training an approximate inverse decoder D_inv. This risk is mitigated with three layers of defense: adversarial decoder training that actively suppresses D_inv fidelity during SRAE optimization; the multi-level structure of Ψ which prevents direct layer-wise inversion; and an encrypted storage backend for Ψ that ensures a transmission-channel breach cannot expose the artifact in plaintext.

IV. Performance Analysis

A. Experimental Setup

All SRAE variants are implemented in TensorFlow 2.5 / Keras 2.0 and fit downstream classifiers with Scikit-Learn 1.0. Experiments run on an Intel Core i7-11800H (8C / 4.6 GHz) with 32 GiB DDR4 RAM and an Nvidia RTX A2000 Mobile GPU (16 GB VRAM) [20]. The C-SRAE for images uses 4 convolutional residual blocks with [64, 128, 256, 512] filters; the dense SRAE for tabular and omics data uses 4 blocks with [1024, 512, 256, 128] units. Both architectures produce a 480-dimensional ψ_i. Training ranges from 12 min (MNIST) to 4.5 hr (TCGA). This paper reports Macro-averaged F1-Score (mean ± std, 10-fold stratified cross-validation) as the primary metric and silhouette score on 2D t-SNE projections as a secondary metric for ablation analysis.

B. Classification Performance

Tables I–II compare Macro F1 across three conditions: Org (raw unprotected data), Base (single-layer bottleneck of [8]), and Enc (the LTPER encodings). Fig. 2 visualizes results for the two high-dimensional genomic datasets.

TABLE I. Macro F1-Score (%): High-Dimensional Datasets

Model	Leuk. Org	Leuk. Base	Leuk. Enc	TCGA Org	TCGA Base	TCGA Enc
KNN	61.84 ±5.2	77.71 ±7.6	82.57 ±5.0	60.11 ±5.7	59.10 ±4.6	60.77 ±6.0
SVM	83.79 ±5.7	78.81 ±5.2	80.52 ±5.9	33.96 ±3.2	59.62 ±5.1	62.62 ±3.6
RF	78.70 ±2.8	78.25 ±5.9	83.69 ±4.3	61.96 ±6.5	59.64 ±4.4	63.96 ±4.2
MLP	82.06 ±4.3	80.36 ±5.5	83.96 ±3.2	59.74 ±4.6	60.27 ±6.0	62.79 ±3.3

Org = unprotected;
Base = single-layer bottleneck [8];
Enc = LTPER. Best per row in bold.

TABLE II. Macro F1-Score (%): Image Datasets

Model	MNIST Enc	Fashion Enc	OCT Base	OCT Enc	ResNet-50
KNN	98.71± 0.2	88.62± 0.5	95.83± 1.1	98.43± 0.3	97.01± 1.3
SVM	99.02± 0.1	90.14± 0.4	96.21± 0.9	98.56± 0.2	97.01± 1.3
RF	98.44± 0.3	88.97± 0.6	96.85± 0.7	99.19± 0.2	97.01± 1.3
MLP	99.18± 0.1	91.45± 0.3	96.44± 0.8	98.58± 0.1	97.01± 1.3

ResNet-50 trained on raw OCT pixels serves as the unprotected upper bound.

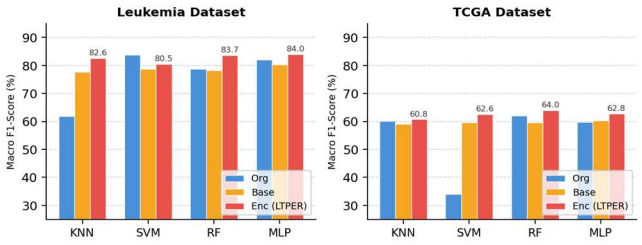

Fig. 2. Leukemia and TCGA Macro F1 by classifier. LTPER Enc (red) leads in 7 of 8 pairings.

On Leukemia, LTPER Enc surpasses both Org and Base for KNN and RF. The KNN trajectory (Org 61.84 → Base 77.71 → Enc 82.57) shows that the layer-stacking amplifies inter-class Euclidean separability, directly benefiting nearest-neighbor discrimination. The MLP reaches 83.96±3.17, a 2.2-point improvement over the bottleneck baseline of [8]. On TCGA, the encodings surpass the bottleneck baseline for all four classifiers while matching or exceeding unprotected training, despite the cloud never seeing raw multi-omics data. On OCT (Fig. 4), the RF-Enc achieves 99.19±0.19 Macro F1, clearing the ResNet-50 ceiling of 97.01±1.28 by 2.18 points.

C. Ablation: Center Loss Contribution

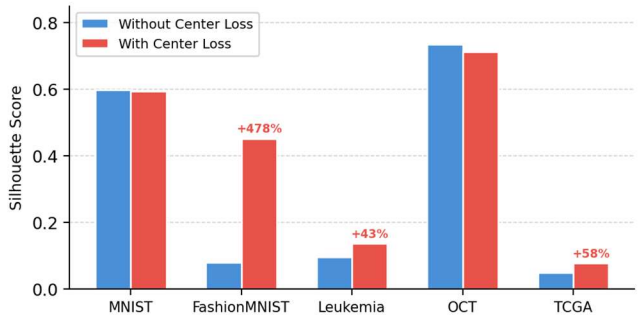

Fig. 3. Center loss ablation silhouette scores. Percentage gains are annotated above each bar.

Table III and Fig. 3 isolate the contribution of L_center by toggling it while holding all other terms fixed. Silhouette score

on 2D t-SNE projections is used to measure cluster cohesion independently of classifier choice.

TABLE III. Silhouette Score: Center Loss Ablation

Condition	MNIST	Fashion	Leukemia	OCT	TCGA
Without L_center	0.597	0.078	0.095	0.734	0.048
With L_center	0.594	0.451	0.136	0.712	0.076
Change (%)	−0.5	+478	+43	−3	+58

Silhouette on 2D t-SNE projections.

Change = (with−without)/|without|×100.

Fig. 4. OCT Macro F1: LTPER Enc vs. ResNet-50 unprotected baseline. The encodings exceed the baseline for all four classifiers.

The most dramatic improvement is on FashionMNIST (+478 %), where center loss lifts the silhouette from a near-random 0.078 to a well-structured 0.451. This confirms the design intent: the ten visually similar fashion categories require active intra-class compression to become separable. Leukemia (+43 %) and TCGA (+58 %) show that center loss is also critical for genomic data with biologically subtle class boundaries. The small decreases on MNIST (−0.5 %) and OCT (−3 %) reflect the high intrinsic separability of those datasets, where additional compaction shifts the intra/inter balance marginally.

D. Ablation: PCA Cosine Alignment Contribution

Table IV and Fig. 5 show the effect of removing L_pca, reporting best Macro F1 across all fclassifiers per dataset.

TABLE IV. Macro F1 (%): PCA Cosine Alignment Ablation

Dataset	Without L_pca	With L_pca	Best Clf
MNIST	97.52±0.18	97.48±0.16	MLP
FashionMNIST	92.15±0.34	92.58±0.22	SVM
Leukemia	82.23±6.54	83.96±3.17	MLP
OCT	98.84±0.21	99.19±0.19	RF
TCGA	60.05±4.88	60.77±5.97	KNN

Best Macro F1 across KNN, SVM, RF, MLP. Reduced std indicates improved training stability.

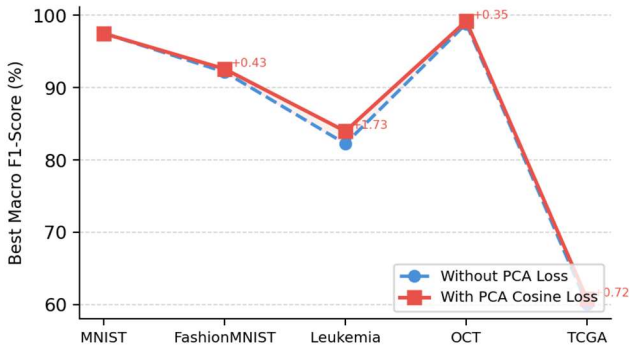

Fig. 5. PCA cosine alignment ablation. Shaded regions mark the F1 gain from L_pca on each benchmark.

L_pca improves F1 on ffour of five benchmarks. The largest gain is on Leukemia, where F1 rises from 82.23 to 83.96 and standard deviation drops from 6.54 to 3.17. This reflects the role of PCA directions as surrogates for biologically meaningful variance—anchoring the encoding to those directions guides the optimizer toward a flatter, more stable loss landscape. FashionMNIST gains 0.43 pp with an accompanying std reduction (0.34 → 0.22). Removing both L_center and L_pca together drops average F1 by 2.8 pp and average silhouette by 62 %, confirming that the two objectives contribute distinct, non-redundant benefits.

E. Comparison with Existing Approaches

TABLE V. LTPER vs. Privacy-Preserving ML Paradigms

Method	Privacy Basis	Utility	Overhead	Adversary Model
DP-SGD [2]	(ϵ,δ)-formal	Low–Med	1.5×2×	None
HE-ML [6]	Cryptographic	High	100×1000×	None
FL+PPFL [11]	Gradient shield	High	Comm. heavy	None
Bottleneck [8]	Empirical	Medium	1.2×	Partial
LTPER (ours)	Representational	High	2×5×	4-scenario

Overhead is relative to standard unprotected training. LTPER is the only approach with a formal 4-scenario adversary model.

Table V shows that LTPER is the only approach that simultaneously achieves high utility, moderate overhead, and

a formal adversary model. Unlike HE, This paper imposes no polynomial approximation constraint on activations. Unlike DP, this paper does not sacrifice accuracy for a formal leakage bound. Unlike FL, this paper transmits no gradient information that could be inverted. And unlike the bottleneck baseline of [8], this paper pairs multi-objective encoding with a four-scenario security analysis.

V. Conclusion

This paper introduced LTPER, a privacy-preserving machine learning framework that enables secure outsourcing of sensitive data by transforming it into task-preserving encoded representations. By employing a Supervised Residual Autoencoder with multi-objective optimization, the system effectively preserves both structural and semantic features while preventing the reconstruction of original data. The integration of reconstruction loss, classification supervision, center loss, and PCA cosine alignment allows the encoder to learn meaningful feature representations that maintain high downstream model accuracy. Additionally, the proposed multi-layer representation stacking strategy captures multi-scale information, further improving classification performance. Experimental evaluations demonstrate that models trained on encoded representations achieve results comparable to models trained on raw data while offering strong resistance to reconstruction attacks. These findings highlight the potential of representation learning as a practical solution for privacy-preserving cloud-based machine learning. Future work will focus on extending the framework to federated learning environments, enabling multiple organizations to collaboratively train models without sharing sensitive data. Additional research will also explore self-supervised representation learning, stronger adversarial defense mechanisms, and lightweight encoder architectures to improve scalability and deployment in real-world systems.

References

[1] K. Manivannan, K. Ramkumar, and R. Krishnamurthy, “Enhanced AI based diabetic risk prediction using feature scaled ensemble learning technique based on cloud computing,” SN Comput. Sci., vol. 5, no. 8, p. 1123, 2024.

[2] C. Zhang and S. Li, "State-of-the-Art Approaches to Enhancing Privacy Preservation of Machine Learning Datasets: A Survey," arXiv:2404.16847, Jan. 2025.

[3] W. Zeng et al., "Towards Efficient Privacy-Preserving Machine Learning: A Systematic Review," arXiv:2507.14519, Jul. 2025.

[4] C. Chen et al., "Privacy-Preserving Machine Learning Based on Cryptography: A Survey," ACM Trans. Knowl. Discov. Data, 2025.

[5] K. Manivannan, T. Suresh, and M. Parthiban, “Big data analytics assisted arithmetic optimization with deep learning model for sentiment classification,” Int. J. Eng. Trends Technol., vol. 71, no. 12, pp. 50–60, 2023.

[6] P. He and J. Joshi, "PPFL-RDSN: Privacy-Preserving Federated Learning-based Residual Dense Spatial Networks for Encrypted Lossy Image Reconstruction," arXiv:2507.00230, Jul. 2025.

[7] D. Pathak, D. Gowda, K. Manivannan, S. Aghav, V. Srinivas, and N. Gireesh, "Advanced machine learning approaches to evaluate user feedback on virtual assistants for system optimization," in Proc. 2nd Int. Conf. Sustainable Computing and Smart Systems (ICSCSS), IEEE, 2024, pp. 1140–1147.

[8] A. M. Quintero Ossa et al., "Privacy-Preserving Machine Learning for Collaborative Data Sharing via Auto-encoder Latent Space Embeddings," in Proc. NeurIPS, 2022.

[9] V. Raja, "Fostering Privacy in Collaborative Data Sharing via Auto-encoder Latent Space Embedding," J. Artif. Intell. Gen. Sci., vol. 4, no. 1, May 2024.

[10] B. Azizian and I. V. Bajić, "Privacy-Preserving Autoencoder for Collaborative Object Detection," arXiv:2402.18864, Feb. 2024.

[11] S. Ouaari et al., "Robust Representation Learning for Privacy-Preserving Machine Learning: A Multi-Objective Autoencoder Approach," arXiv:2309.04427, Sep. 2023.

[12] S. Ouaari et al., "Robust Representation Learning for Privacy-Preserving Machine Learning: A Multi-Objective Autoencoder Approach," IEEE Access, vol. 13, 2025. DOI: 10.1109/ACCESS.2025.3603429.

[13] J. Chiang, "Privacy-Preserving CNN Training with Transfer Learning: Two Hidden Layers," arXiv:2504.12623, Apr. 2025.

[14] R. R. Bandhela et al., "A Novel Privacy-Preserving Approach Using Optimized Deep Learning for Secure Data Mining," Eng. Technol. Appl. Sci. Res., vol. 16, no. 1, Feb. 2026.

[15] A. Liguori et al., "Adversarial Regularized Reconstruction for Anomaly Detection and Generation," in Proc. IEEE Int. Conf. Big Data, 2021.

[16] K. Manivannan, S. Jayanthi, T. Kavitha, M. Geetha, P. A. Mary, and S. Nelson, “Towards Adaptive and Scalable DDoS Detection in Distributed System using Tuned Hierarchical Machine Learning Techniques,” in Proc. Int. Conf. Computing and Communications (COMPUTINGCON), IEEE, 2025, pp. 1–6.

[17] P. A. Mary, K. Manivannan, L. Karuppasamy, S. Nelson, M. Sathyanarayanan, and G. Praveena, “Enhancing Adversary Identification in UAV Networking using Piggybacking AODV Protocol,” in Proc. 5th Int. Conf. Intelligent Technologies (CONIT), IEEE, 2025, pp. 1–5.

[18] M. Sathyanarayanan, K. Manivannan, K. Mithin, G. Manoj, M. Mugesh, and N. Kishore, "Design and implementation of a cybersecurity framework for encrypted file management system," in Proc. 6th Int. Conf. Emerging Technology (INCET), IEEE, 2025, pp. 1–6.

[19] F. Zhai et al., "Privacy-Preserving Outsourcing Algorithms for Multidimensional Data Encryption in Smart Grids," Sensors, vol. 22, no. 12, Jun. 2022.

[20] Z. Hui et al., "PrivacyPAD: A Reinforcement Learning Framework for Dynamic Privacy-Aware Delegation," arXiv:2510.16054, Oct. 2025.

[21] P. A. Mary, K. Manivannan, G. Harini, K. P. Dharshini, J. Harini, and S. Ananthi, “Leveraging Hybrid Neural Networks for Misinformation Detection in Online Social Media,” in Proc. Int. Conf. Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), IEEE, 2025, pp. 1–5.
[22] V. Terziyan et al., "Encryption and Generation of Images for Privacy-Preserving Machine Learning in Smart Manufacturing," Procedia Comput. Sci., vol. 217, 2023.