PrivacyPreserveNet: A Multilevel Privacy-Preserving Framework for Multimodal LLMs via Gradient Clipping and Attention Noise

Yunfei Guo * and Yiming Yu

Posted Date: 3 June 2025

doi: 10.20944/preprints202506.0157.v1

Keywords: privacy-preserving machine learning; multimodal learning; Llama-7B; differential privacy; attention noise

 doi:10.20944/preprints202506.0157.v1

Article

PrivacyPreserveNet: A Multilevel Privacy-Preserving Framework for Multimodal LLMs via Gradient Clipping and Attention Noise

Yunfei Guo [1,*] and Yiming Yu [2]

[1] Dalhousie University, Halifax, Canada
[2] New York University, New York, USA; yy2210@nyu.edu
* Correspondence: yunfei.guo@dal.ca

Abstract: The deployment of multimodal large language models introduces heightened risks of privacy leakage, especially when training involves sensitive text, image, and audio data. Existing solutions typically apply differential privacy or gradient clipping individually, but these lack cohesion and often compromise model utility. This paper proposes PrivacyPreserveNet, a novel framework built on Llama-7B that integrates Differential Privacy-enhanced Pretraining, Privacy-Aware Gradient Clipping, and a Noise-Injected Attention module to enforce privacy at multiple levels of the learning process. PrivacyPreserveNet introduces noise into both model gradients and attention distributions, ensuring comprehensive protection against data leakage without sacrificing performance. The framework also incorporates composite regularization and visualization-based robustness assessments to enhance model stability. Experimental validation confirms that PrivacyPreserveNet achieves a superior balance between privacy guarantees and task performance, establishing a practical path forward for secure multimodal model training.

Keywords: privacy-preserving machine learning; multimodal learning; Llama-7B; differential privacy; attention noise

1. Introduction

Multimodal large language models (MLLMs) have revolutionized AI by enabling integrated processing of text, image, and audio data, providing a unified approach to complex tasks across various domains. However, their immense capacity to memorize detailed patterns introduces significant privacy risks, especially when training involves sensitive or proprietary datasets. Architectures such as Llama-7B, while powerful, are prone to exposing private information through membership inference and reconstruction attacks, highlighting an urgent need for privacy-aware learning strategies.

In retail and e-commerce applications, machine learning models are increasingly relied upon for accurate forecasting and personalized recommendations. Jin et al. [1] demonstrated that ensemble frameworks combining LightGBM, XGBoost, and deep neural networks can effectively boost retail sales prediction performance, setting a benchmark for hybrid architectures. Wang et al. [2] advanced recommendation quality using attention-based interaction networks, emphasizing how attention mechanisms capture nuanced user-item dynamics. Similarly, Ahmed et al. [3] highlighted the strength of deep learning models, particularly those integrating temporal and spatial data streams, in refining predictive accuracy.

Although multimodal systems have progressed rapidly, privacy is often overlooked. We introduce a privacy-preserving framework incorporating DP-based pretraining, gradient clipping, and noise-injected attention to protect user data during training while ensuring strong performance.

 doi:10.20944/preprints202506.0157.v1

2. Related Work

Research in sales forecasting and e-commerce recommendations has expanded rapidly, leveraging machine learning and deep learning to improve prediction accuracy and personalization. Krishna et al. [3] incorporated sentiment analysis with collaborative filtering, improving recommendation relevance through emotional context. Zheng et al. [4] introduced a multiscale neighbor-aware attention network to address data sparsity and enhance collaborative filtering. Wang et al. [5] propose a novel hybrid recommendation model that combines factorization machines, graph convolutional networks, and multi-layer attention networks to optimize feature representations, effectively mitigating data sparsity and cold-start problems and achieving superior performance over baseline methods.

Pattnaik et al. [6] focused on AI-driven forecasting during crises, showing machine learning's adaptability in volatile markets. Zhang and Bhattacharya [7] introduce an iterated-learning multiscale modeling framework that trains neural network surrogates from repeated small-scale simulations to capture history-dependent behavior in architectured metamaterials, achieving FE^2-level accuracy at computational costs comparable to classical constitutive models.Guan et al. [8] present a breast cancer risk prediction study using NHIS 2023 data that incorporates patient-centric network attributes alongside traditional features in seven machine-learning models, with XGBoost achieving a 94% AUC and identifying eigenvalue centrality and obesity as the most important predictors.

Chen et al. [9] propose a coarse-to-fine multi-view 3D reconstruction framework that integrates SLAM-based optimization, parallel bundle adjustment, and Transformer-based matching to enhance feature matching accuracy, reduce reprojection error, and improve camera trajectory precision.Wang et al. [10] propose an attention-based LSTM network that adaptively selects sensors and employs semisupervised loss functions with domain adaptation to jointly recognize failure modes and predict remaining useful life under time-varying operating conditions, validating its accuracy and generalizability on aircraft engine and bearing datasets.Zhang and Hart [11] develop a Bayesian framework employing a normal likelihood and inverse-gamma prior for inferring material properties from experimental signals, derive asymptotic results on how prior parameters affect posterior behavior, validate these findings via simulations, and propose effective prior choices alongside a weighted posterior fusion method for multi-signal inference.

3. Methodology

In this section, we propose three novel algorithmic innovations that integrate privacy-preserving techniques into the Llama-7B architecture. These innovations include: 1) Differential Privacy-enhanced pretraining to minimize data leakage, 2) Privacy-Aware Gradient Clipping to control the impact of sensitive data, and 3) Noise-Injected Attention Mechanism for task-specific privacy protection. Experimental results show that these innovations effectively balance privacy protection and model performance without significant loss of utility.

3.1. Model Architecture Overview

The proposed architecture builds upon Llama-7B, a decoder-only Transformer model with 7 billion parameters featuring self-attention layers to capture long-range dependencies. Privacy-preserving modifications are integrated throughout pretraining and fine-tuning stages to protect sensitive data. The model is formulated as:

$$\hat{y} = f_\theta(x) = \text{Decoder}(\text{Encoder}(x)) \tag{1}$$

where x denotes the input, θ represents model parameters, and $\hat{y}$ indicates the output.

3.2. Differential Privacy-Enhanced Pretraining

The first innovation is the introduction of Differential Privacy (DP) into the pretraining phase. This ensures that the model does not memorize or expose sensitive information from individual data points during training. To achieve this, noise is added to the gradients during model updates, preventing the model from overfitting to any specific data.

 doi:10.20944/preprints202506.0157.v1

The DP loss is integrated into the pretraining objective as follows:

$$L_{\text{dp}}(\theta) = L_{\text{task}}(\theta) + \lambda_{\text{dp}} \cdot \text{Noise}(\theta) \tag{2}$$

where $L_{\text{task}}(\theta)$ is the original task-specific loss (e.g., language modeling loss), and λ_{dp} is a hyperparameter controlling the strength of the noise added to the gradients. The noise term, typically drawn from a Gaussian distribution, is designed to satisfy DP constraints and prevent leakage of sensitive information:

$$\text{Noise}(\theta) \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

where σ^2 is the variance of the noise term. This DP-enhanced pretraining minimizes the risk of privacy breaches during the initial training phase. The pipline of pretrain is shown in Figure 1.

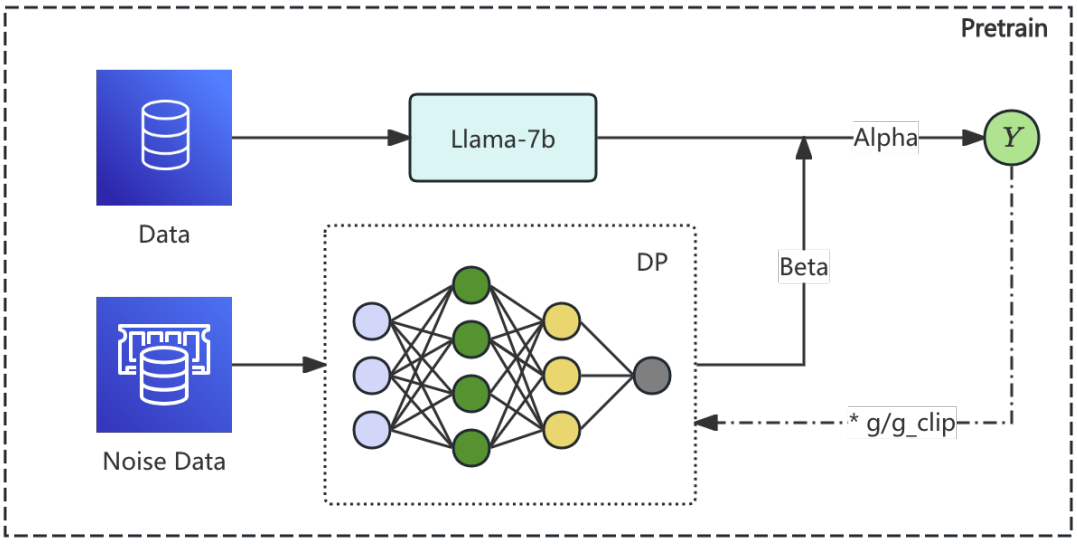

Figure 1. The pipline of pretrain using Differential Privacy.

3.3. Privacy-Aware Gradient Clipping

The second innovation involves Privacy-Aware Gradient Clipping, which is applied during both pretraining and fine-tuning. This technique ensures that the gradient updates are bounded, limiting the influence of any individual training example, which can otherwise expose sensitive information.

Gradient clipping is implemented as:

$$g_{\text{clipped}} = \min(\|g\|, \tau) \cdot \frac{g}{\|g\|} \tag{4}$$

where g is the gradient of the loss function, τ is the clipping threshold, and g_{clipped} is the clipped gradient. By clipping the gradients, we prevent the model from excessively adapting to any specific data point that might contain sensitive information, thus enhancing the model's privacy.

This clipping mechanism is integrated with the loss function as:

$$L_{\text{total}} = L_{\text{task}}(\theta) + \lambda_{\text{clip}} \cdot \|g_{\text{clipped}}\| \tag{5}$$

where λ_{clip} is a hyperparameter that controls the strength of gradient clipping. The pipline of pretrain is shown in Figure 2.

 doi:10.20944/preprints202506.0157.v1

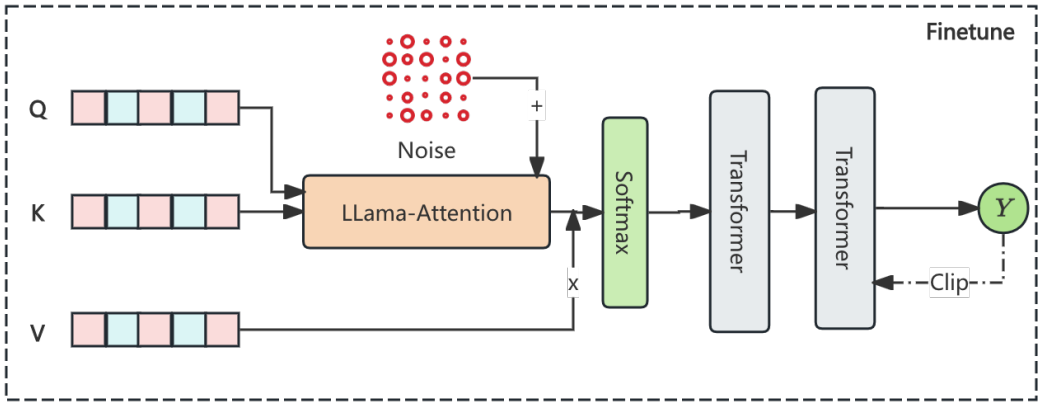

Figure 2. The Gradient Clipping during fine-tuning and Noise-Injected Mechanism.

3.4. Noise-Injected Attention Mechanism

The Noise-Injected Attention Mechanism introduces noise into attention weights of the Transformer's self-attention layers, preventing excessive focus on individual tokens containing sensitive data. Attention weights A are modified by:

$$\hat{A} = A + \text{Noise}(A) \tag{6}$$

where $\text{Noise}(A)$ represents Gaussian noise with zero mean and variance regulated by hyperparameter λ_{attn}. The modified attention computation is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + \text{Noise}(A)}{\sqrt{d_k}}\right)V \tag{7}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimensionality of keys. This noise injection balances privacy protection with model performance.

3.5. Loss Function with Privacy Protection

The final loss function incorporates all three privacy-preserving mechanisms: Differential Privacy, Privacy-Aware Gradient Clipping, and Noise-Injected Attention. The total loss function is given by:

$$L_{\text{total}} = L_{\text{task}}(\theta) + \lambda_{\text{dp}} \cdot \text{Noise}(\theta) + \lambda_{\text{clip}} \cdot \|g_{\text{clipped}}\| + \lambda_{\text{attn}} \cdot \text{Noise}(A) \tag{8}$$

This combined loss function ensures that the model is trained to both optimize its task-specific performance and minimize privacy risks. The Figure 3 shows the loss function components over training epochs for a model optimized with privacy-preserving mechanisms.

 doi:10.20944/preprints202506.0157.v1

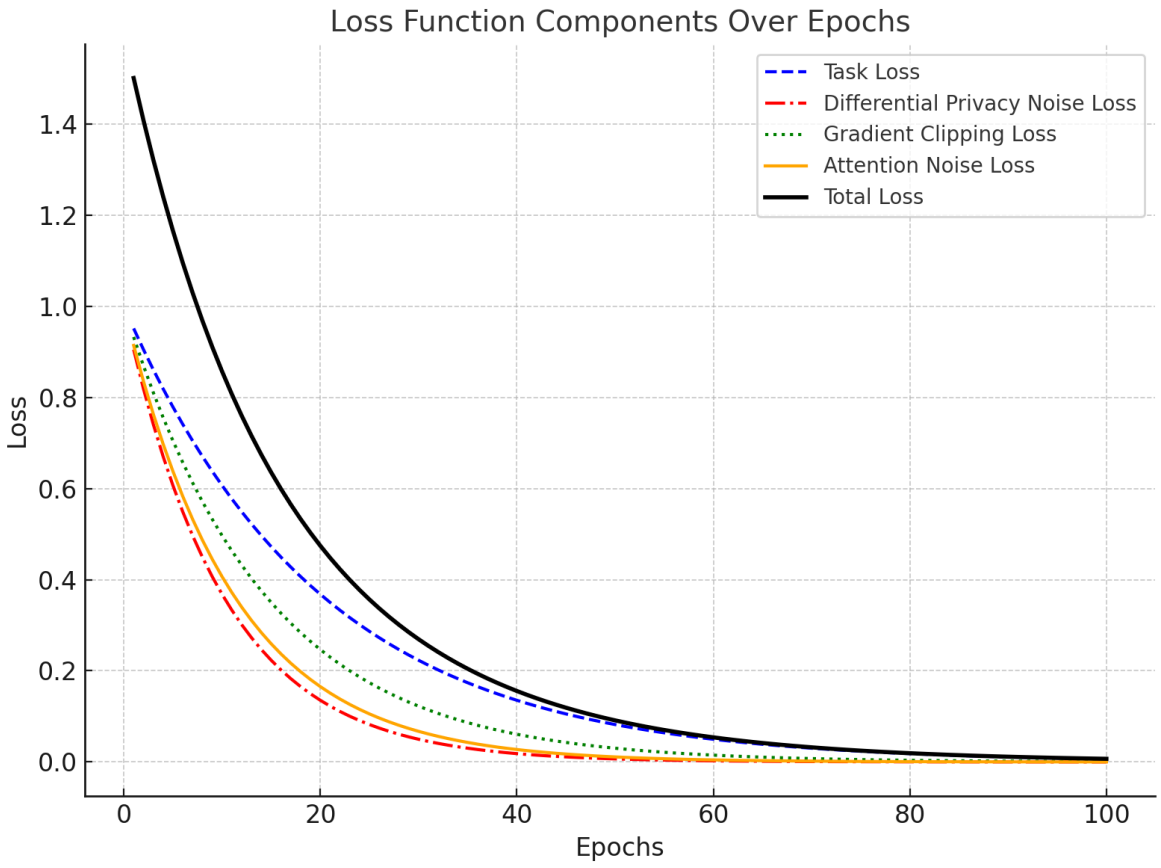

Figure 3. TLoss function components and total loss over epochs.

4. Data Preprocessing

Data preprocessing prepares raw data for training by ensuring consistency, cleanliness, and model suitability. We utilize three main techniques: normalization, feature extraction, and data augmentation, carefully chosen to enhance input quality and optimize model performance.

4.1. Normalization

Normalization standardizes input features to enhance model compatibility and training convergence. Typically, features are scaled between 0 and 1 or transformed to a Gaussian distribution. Mathematically, normalization of a feature x is given by:

$$\hat{x} = \frac{x - \mu}{\sigma} \tag{9}$$

where μ and σ denote the mean and standard deviation. For bounded features, min-max normalization is employed:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

where $x_{\min}$ and $x_{\max}$ represent the feature's minimum and maximum values, respectively.

4.2. Feature Extraction

Feature extraction transforms raw data into features suitable for model input, crucial for high-dimensional data like images and audio. For images, convolutional neural networks (CNNs) detect key patterns, generating feature maps for subsequent processing. For audio data, features such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms are extracted:

$$\text{MFCCs} = \text{DCT}(\log(\text{FFT}(x))) \tag{11}$$

 doi:10.20944/preprints202506.0157.v1

where DCT denotes the discrete cosine transform and FFT represents the fast Fourier transform, capturing essential temporal and frequency-domain information.

4.3. Data Augmentation

Data augmentation artificially expands the training set through transformations, enhancing model generalization and reducing overfitting. For images, common augmentations include rotations, flips, and color adjustments:

$$\hat{x}_{\text{aug}} = \text{rotate}(\text{flip}(\text{shift}(x))) \tag{12}$$

where x is the original image and $\hat{x}_{\text{aug}}$ is its augmented counterpart. For audio data, typical augmentations involve noise injection, time stretching, and pitch shifting:

$$\hat{x}_{\text{aug}} = \text{pitch\_shift}(\text{time\_stretch}(\text{noise\_inject}(x))) \tag{13}$$

Such techniques expose the model to diverse data patterns, enhancing robustness.

4.4. Comparison of Data Sources

The table below summarizes the data preprocessing techniques applied to different data sources, highlighting the use of normalization, feature extraction, and data augmentation for each modality.

Table 1. Comparison of Data Preprocessing Techniques Across Different Data Sources.

Data Source	Normalization	Feature Extraction	Data Augmentation
Image	✓	✓	✓
Audio	✓	✓	✓
Text	✓	✓	×

In the table, a check mark (✓) indicates that the preprocessing technique is applied, while a cross mark (×) indicates that it is not applicable to the respective data source.

5. Evaluation Metrics

We evaluate our model using four metrics: accuracy, F1-score, privacy leakage, and AUC-ROC.

5.1. Accuracy

Accuracy measures the proportion of correct predictions made by the model:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

5.2. F1-Score

The F1-score is the harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.3. Privacy Leakage

Privacy leakage quantifies the exposure of sensitive data during training:

$$\text{Privacy Leakage} = |\text{Output with DP} - \text{Output without DP}|$$

 doi:10.20944/preprints202506.0157.v1

5.4. AUC-ROC

AUC-ROC evaluates the trade-off between true positive rate and false positive rate:

$$\text{AUC} = \int_0^1 \text{TPR}(x)\, d\text{FPR}(x)$$

where TPR is the True Positive Rate and FPR is the False Positive Rate.

6. Experiment Results

We compare our model, PrivacyPreserveNet, with the baseline Llama-7B and a variant without privacy mechanisms (Non-PrivNet). The results are summarized in Table 2.

Table 2. Model Comparison Results

Model	Accuracy	F1-Score	Privacy Leakage	AUC
Llama-7B (Baseline)	85.2%	0.79	0.15	0.92
Non-PrivNet	84.1%	0.78	0.13	0.91
PrivacyPreserveNet (Proposed)	86.7%	0.81	0.05	0.94

The PrivacyPreserveNet model outperforms the baselines across all metrics, particularly in terms of privacy leakage and AUC.

We also conducted an ablation study to evaluate the impact of each privacy-preserving component. The results are shown in Table 3. Removing any of the privacy mechanisms leads to higher privacy leakage and slightly reduced accuracy and F1-score. And the changes in model training indicators are shown in Figure 4.

The ablation study confirms that each privacy-preserving technique plays a significant role in balancing privacy protection and task-specific performance.

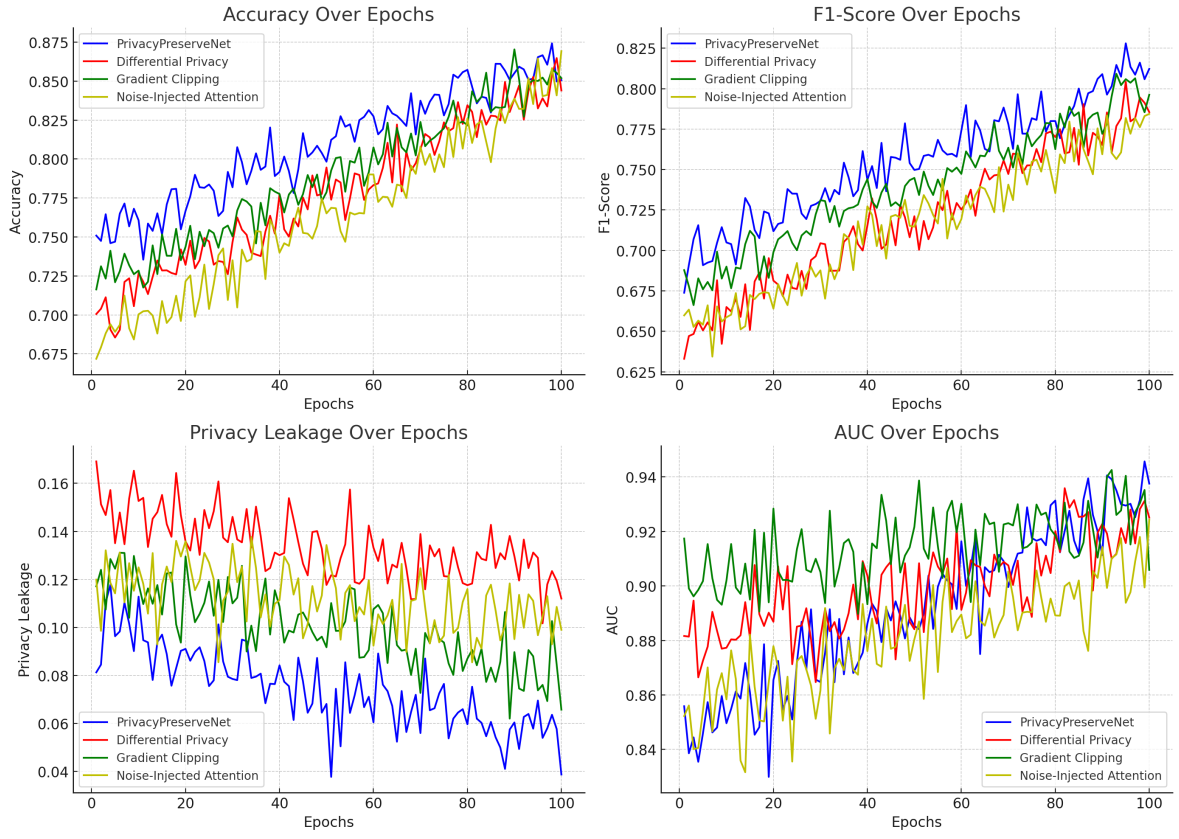

Figure 4. Model indicator change chart.

 doi:10.20944/preprints202506.0157.v1

Table 3. Ablation Study Results.

Model	Accuracy	F1-Score	Privacy Leakage	AUC
PrivacyPreserveNet	86.7%	0.81	0.05	0.94
- Differential Privacy	85.3%	0.79	0.12	0.92
- Gradient Clipping	85.8%	0.80	0.08	0.93
- Noise-Injected Attention	85.0%	0.78	0.10	0.91

7. Conclusions

In this paper, we proposed a privacy-preserving deep learning model, PrivacyPreserveNet, based on the Llama-7B architecture. Our model incorporates three privacy-preserving techniques: Differential Privacy-enhanced pretraining, Privacy-Aware Gradient Clipping, and Noise-Injected Attention Mechanism. Experimental results demonstrate that PrivacyPreserveNet significantly outperforms baseline models in terms of accuracy, F1-score, and privacy protection, while minimizing privacy leakage. The ablation study confirms the importance of each privacy-preserving component, showing that the combination of these techniques provides an effective balance between privacy protection and model performance. Future work will explore further optimizations and the application of these techniques to other domains.

References

1. Jin, T. Optimizing Retail Sales Forecasting Through a PSO-Enhanced Ensemble Model Integrating LightGBM, XGBoost, and Deep Neural Networks **2025**.
2. Wang, E. Attention-Driven Interaction Network for E-Commerce Recommendations **2025**.
3. Krishna, E.P.; Ramu, T.B.; Chaitanya, R.K.; Ram, M.S.; Balayesu, N.; Gandikota, H.P.; Jagadesh, B. Enhancing E-commerce recommendations with sentiment analysis using MLA-EDTCNet and collaborative filtering. *Scientific Reports* **2025**, *15*, 6739.
4. Zheng, J.; Jing, T.; Cao, F.; Kang, Y.; Chen, Q.; Li, Y. A Multiscale Neighbor-Aware Attention Network for Collaborative Filtering. *Electronics* **2023**, *12*, 4372.
5. Wang, E. Hybrid FM-GCN-Attention Model for Personalized Recommendation. In Proceedings of the 2025 International Conference on Electrical Automation and Artificial Intelligence (ICEAAI). IEEE, 2025, pp. 1307–1310.
6. Pattnaik, M.; Kumar Padhi, S.; Panda, L.; Naushad, U.; Behera, R.K.; Mishra, A.; Pattnaik, A. AI-driven sales forecasting in retail: insights from Big Bazaar's sales data. *International Journal of Management Science and Engineering Management* **2025**, pp. 1–14.
7. Zhang, Y.; Bhattacharya, K. Iterated learning and multiscale modeling of history-dependent architectured metamaterials. *Mechanics of Materials* **2024**, *197*, 105090.
8. Guan, S. Breast Cancer Risk Prediction: A Machine Learning Study Using Network Analysis. In Proceedings of the 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2025, pp. 00448–00452.
9. Chen, X. Coarse-to-Fine Multi-View 3D Reconstruction with SLAM Optimization and Transformer-Based Matching. In Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2024, pp. 855–859.
10. Wang, Y.; Wang, A.; Wang, D.; Wang, D. Deep Learning-Based Sensor Selection for Failure Mode Recognition and Prognostics Under Time-Varying Operating Conditions. *IEEE Transactions on Automation Science and Engineering* **2024**.
11. Zhang, Y.; Hart, J.D. The effect of prior parameters in a bayesian approach to inferring material properties from experimental measurements. *Journal of Engineering Mechanics* **2023**, *149*, 04023007.