# Evaluating the Effectiveness of Posterior Truncation in Mitigating Membership Inference Attacks on Public Image Classification Models

Jenom John Gimba
School of Computing
*National College of Ireland*
Dublin, Ireland
x23373750@student.ncirl.ie

***Abstract*— Membership inference attacks (MIAs) pose significant privacy risks to machine learning models by determining whether specific data points were used in training. In this work, I evaluate the effectiveness of posterior truncation as a defense against membership inference attacks on public image classification models. I utilize models from the SecurityNet database and create controlled variants with different overfitting levels using CIFAR-10 and CIFAR-100 datasets. My evaluation encompasses three attack methodologies: Prediction Correctness Attack (PCA), Modified Prediction Entropy (MPE), and MLP-based attacks, tested against multiple truncation strategies. I find that posterior truncation significantly reduces attack effectiveness, with top-1 truncation achieving up to 0.232 AUC reduction against entropy-based attacks (p=0.016, Cohen's d=2.83). My analysis reveals attack-specific vulnerabilities: MPE attacks are highly susceptible to truncation, while PCA attacks remain immune due to their reliance on prediction correctness rather than posterior distributions. These findings provide practical guidance for deploying privacy-preserving defenses in real-world machine learning systems while maintaining computational efficiency.**

***Keywords—Posterior truncation, securitynet, membership interference attacks, cifar***

## I. Introduction

Machine learning models have achieved remarkable success across diverse domains, from healthcare diagnostics to autonomous systems. However, their widespread deployment has raised significant privacy concerns, particularly regarding the potential leakage of sensitive information about training data. Membership inference attacks (MIAs) represent a critical privacy threat where adversaries attempt to determine whether specific data points were used to train a target model [2]. These attacks exploit the tendency of machine learning models to exhibit different prediction behaviors on training data versus unseen data, potentially exposing sensitive information about individuals whose data was used for training.

The privacy implications of membership inference attacks are particularly concerning in sensitive domains such as healthcare, finance, and personal data analytics. For instance, an adversary capable of determining whether a patient's medical record was used to train a disease prediction model could infer sensitive health information about that individual [3]. Similarly, in financial applications, successful membership inference could reveal participation in specific programs or services, violating user privacy expectations [4].

Recent research has demonstrated that membership inference vulnerabilities are not merely theoretical concerns but represent practical threats against real-world systems [2], [5]. These attacks have proven effective against various model architectures and have been successfully demonstrated against commercial machine learning services [2]. The relationship between model overfitting and membership inference vulnerability has been theoretically and empirically established, with overfitted models exhibiting higher susceptibility to such attacks [6], [7].

To address these privacy risks, researchers have proposed various defense mechanisms, including differential privacy [8], knowledge distillation [9], and regularization techniques [10]. However, most existing defenses either significantly degrade model utility or require substantial computational overhead, limiting their practical applicability. Moreover, the majority of privacy research has focused on models trained specifically for research purposes, raising questions about the effectiveness of proposed defenses on realistic, high-quality models deployed in practice.

The SecurityNet database [1] addresses this limitation by providing a comprehensive collection of 910 publicly available image classification models trained for various purposes, offering a more realistic evaluation platform for security and privacy research. Initial analysis using SecurityNet has confirmed that membership inference attack effectiveness correlates strongly with model overfitting levels, validating theoretical predictions on practical systems [1]. However, the evaluation of defense mechanisms against membership inference attacks on such realistic models remains underexplored.

Posterior truncation represents a promising yet understudied defense approach that involves limiting the information revealed in model outputs by retaining only the most confident predictions. Unlike complex defenses that require model retraining or architectural modifications, posterior truncation can be applied as a post-processing step to existing deployed models. This approach is motivated by the observation that membership inference attacks often exploit fine-grained differences in prediction confidence that may not be essential for the model's primary task performance.

Despite its intuitive appeal and practical advantages, posterior truncation has received limited systematic evaluation, particularly on realistic pretrained models. Existing studies have primarily focused on researcher-trained models with controlled settings, leaving significant gaps in understanding how this defense performs across different attack methodologies, model architectures, and overfitting levels in practical deployment scenarios.

### Research Question and Objectives

This research addresses the fundamental question: *How does posterior truncation affect the effectiveness of membership inference attacks on public image classification models?* To answer this question, I establish several specific objectives that guide the investigation.

My primary objective is to systematically evaluate the effectiveness of posterior truncation strategies across multiple membership inference attack methodologies. I aim to determine which truncation approaches provide optimal privacy-utility trade-offs and identify the conditions under which these defenses are most effective. Additionally, I seek

to validate the correlation between model overfitting and membership inference vulnerability using controlled, realistic models.

I also investigate attack-specific vulnerabilities to understand how different membership inference methodologies respond to truncation defenses. This analysis provides insights into the underlying mechanisms that make certain attacks more susceptible to output perturbation than others. Furthermore, I examine the computational overhead introduced by truncation strategies to assess their practical feasibility for real-world deployment.

In this work, I conduct the first comprehensive evaluation of posterior truncation as a defense against membership inference attacks on public image classification models. I leverage the SecurityNet framework [1] and create controlled variants with different overfitting characteristics using CIFAR-10 and CIFAR-100 datasets. My evaluation encompasses three prominent attack methodologies: Prediction Correctness Attack (PCA), Modified Prediction Entropy (MPE), and MLP-based attacks, tested against multiple truncation strategies including top-k and confidence-based approaches.

The key contributions of this research are validating the correlation between model overfitting and membership inference vulnerability on a controlled set of realistic models, confirming theoretical predictions with empirical evidence. Demonstrate that posterior truncation provides significant defense against entropy-based membership inference attacks, achieving up to 0.232 AUC reduction with statistical significance. Identify attack-specific defense patterns, revealing that different attack methodologies exhibit varying susceptibility to truncation defenses. Finally, show that simple truncation strategies achieve strong privacy protection with minimal computational overhead, making them practical for real-world deployment.

The remainder of this paper is organized as follows. Section II reviews related work on membership inference attacks, privacy-preserving defenses, and public model security evaluation. Section III presents the methodology, including dataset selection, model preparation, attack implementations, and truncation strategies. Section IV provides comprehensive experimental results, statistical analysis, and performance evaluation. Section V discusses the implications of findings, limitations, and directions for future research. Finally, Section VI concludes the paper.

## II. Related Work

### A. Membership Inference Attacks

Membership inference attacks were first formalized by Shokri et al. [2], who demonstrated that machine learning models leak information about their training data through prediction outputs. Their seminal work established the foundation for black-box membership inference by training shadow models to distinguish between member and non-member prediction patterns. This attack methodology has since been refined and extended across various domains and model architectures.

Carlini et al. [7] provided a theoretical foundation for membership inference attacks, demonstrating that these attacks fundamentally exploit the confidence gap between training and test data predictions. Their analysis revealed that attack success correlates strongly with model overfitting, providing theoretical justification for empirical observations. Recent work by Chen et al. [5] has improved attack practicality through difficulty calibration techniques, achieving more consistent performance across different datasets and model configurations.

The scope of membership inference attacks has expanded beyond traditional supervised learning. Li and Zhang [17] demonstrated effective attacks in label-only scenarios where adversaries access only predicted labels without confidence scores. Wang et al. [13] revealed vulnerabilities during the pre-training phase, showing how adversaries can manipulate foundation models to enhance subsequent membership inference attacks. In federated learning contexts, Gomes et al. [14] introduced active attribute inference attacks that exploit gradient updates to infer sensitive attributes from participant data.

Advanced attack methodologies have emerged to exploit specific model behaviors. Liu et al. [15] proposed attacks based on loss trajectory analysis, leveraging the observation that training samples exhibit distinct loss patterns during model training. Leino and Fredrikson [16] developed white-box attacks that exploit model memorization patterns, achieving higher accuracy by directly analyzing model parameters rather than relying solely on predictions.

### B. Privacy-Preserving Defenses

Defense mechanisms against membership inference attacks have evolved along multiple research directions, each offering different privacy-utility trade-offs. Differential privacy has emerged as a prominent approach, with Abadi et al. [8] demonstrating its application to deep learning through differentially private stochastic gradient descent. This approach provides theoretical privacy guarantees but often requires careful hyperparameter tuning to balance privacy and model utility [10].

Knowledge distillation and regularization techniques offer alternative defense strategies. Wang et al. [9] proposed RelaxLoss, a regularization approach that reduces overfitting while maintaining model performance. However, these techniques typically require modifications to the training process, limiting their applicability to existing deployed models.

Recent defense research has focused on addressing practical deployment constraints. Shang et al. [11] specifically examined defenses for iteratively pruned networks, recognizing that model compression techniques may introduce new vulnerabilities. Wu et al. [12] investigated membership inference risks in transfer learning scenarios, highlighting the need for defense mechanisms that account for pre-training and fine-tuning phases.

The effectiveness of existing defenses has been questioned by comprehensive evaluation studies. Song and Mittal [18] conducted systematic privacy risk assessments, revealing that many proposed defenses fail under stronger attack models or provide insufficient protection in practice. This evaluation gap has motivated research toward more robust and practical defense mechanisms.

### C. Public Model Security Evaluation

Traditional security and privacy research has primarily focused on models trained specifically for research purposes, potentially limiting the generalizability of findings to real-world deployments. The SecurityNet database [1] addresses

this limitation by providing 910 publicly available image classification models with comprehensive metadata, enabling more realistic security evaluations.

Zhang et al. [1] demonstrated that attack effectiveness varies significantly between researcher-trained models and public models, with public models often exhibiting different vulnerability patterns due to diverse training objectives and optimization procedures. Their analysis confirmed the correlation between overfitting and membership inference vulnerability across a large-scale model collection, validating theoretical predictions on practical systems.

The SecurityNet framework has enabled new insights into model security patterns across different architectures and datasets. The comprehensive evaluation revealed that attack success rates vary considerably based on model provenance, with benchmark models often showing different vulnerability profiles compared to security-focused models. This finding emphasizes the importance of evaluating privacy defenses on realistic model distributions rather than controlled research settings.

### *D. Output Perturbation and Truncation Techniques*

Output perturbation represents a class of defenses that modify model predictions to reduce information leakage while preserving utility for the primary task. These approaches are particularly attractive for practical deployment because they can be applied as post-processing steps to existing models without requiring retraining.

Traditional output perturbation techniques include adding calibrated noise to predictions or applying temperature scaling to confidence scores. However, these approaches often struggle to balance privacy protection with prediction utility, particularly for tasks requiring high-confidence outputs.

Posterior truncation, the focus of this work, represents a specific form of output perturbation that selectively retains high-confidence predictions while suppressing low-confidence outputs. This approach is motivated by the observation that membership inference attacks often exploit subtle confidence differences that may not be essential for downstream applications.

Despite its intuitive appeal, posterior truncation has received limited systematic evaluation in the literature. Existing studies have primarily focused on theoretical analysis or evaluation on researcher-trained models with controlled overfitting levels. The effectiveness of truncation strategies across different attack methodologies, realistic model architectures, and varying overfitting characteristics remains largely unexplored.

The gap in posterior truncation evaluation is particularly notable given its practical advantages. Unlike differential privacy or regularization techniques, truncation can be implemented without access to training data or model internals, making it applicable to black-box deployed systems. Truncation strategies can be dynamically adjusted based on privacy requirements without model redeployment.

This work addresses the evaluation gap by providing the assessment of posterior truncation effectiveness against multiple attack methodologies on realistic public models. By leveraging the SecurityNet framework [1] and controlled overfitting variants, I systematically evaluate how truncation strategies perform across different threat models.

## III. METHODOLOGY

### *A. Dataset and Model Selection*

This study utilizes a curated subset of the SECURITYNET database [1], which contains 910 publicly available image classification models trained on 42 datasets from 13 categories. To ensure diversity in both dataset complexity and architectural design, two datasets were selected: CIFAR-10, and CIFAR-100. These datasets span low- to high-complexity visual recognition tasks. The CIFAR datasets offer small-scale, coarse- and fine-grained classification challenges

From the SECURITYNET repository, three representative architectures were chosen:

i. ResNet-50: a widely used residual network with strong baseline performance.

ii. VGG-16: a deep convolutional model without skip connections, representing traditional CNN design.

iii. DLA-169: a modern architecture leveraging hierarchical feature aggregation for improved representation learning.

Selection was based on availability of pretrained weights, diversity of design philosophy, and prevalence in both benchmark and security-model subsets of SECURITYNET. Metadata such as parameter counts, FLOPs, dropout usage, and training dataset information were recorded to support reproducibility.

### *B. Data Pre-Processing*

Images were resized to match the expected input size of each model (32×32 for CIFAR datasets, 224×224 for ImageNet-1k) and normalized using the dataset-specific mean and standard deviation provided in SECURITYNET annotations.

For membership inference attack evaluation, datasets were partitioned into:

i. Member set: images used in the model's training data.

ii. Non-member set: images from the original test split, disjoint from the member set.

iii. Development set: Constructed from the non-member pool for hyperparameter tuning of attack models.

To ensure statistical robustness, I maintained a balanced 1:1 ratio between member and non-member samples, with each partition containing at least 1,000 samples per class when available.

### *C. Membership Inference Attack Implementations*

Five attack methodologies were implemented, covering both metric-based and model-based approaches:

Baseline Attacks:

1. **Prediction Correctness Attack (PCA)**: Determines membership by checking if the target model's predicted label matches the ground truth. Members generally exhibit higher correctness rates due to overfitting.

2. **Modified Prediction Entropy (MPE)**: Measures the entropy of posterior distributions. Lower entropy typically indicates a higher likelihood of membership. Entropy is computed as $H(p) = -\sum_i p_i \log p_i$ over the posterior probabilities.
3. **MLP-Based Attack**: Trains a binary classifier (multi-layer perceptron with hidden layers [64, 32]) on the posteriors of the target model for known member and non-member samples. The trained attack model predicts membership for unseen queries.

Advanced Attack Variants:

4. **Adaptive MLP Attack**: Enhanced MLP attack that trains on posteriors processed with various truncation levels to simulate defense-aware scenarios. Shadow models are trained with truncation probabilities sampled from [0.0, 0.3, 0.5, 0.7].
5. **Ensemble Attack:** Combines predictions from multiple attack methods using weighted voting based on individual attack confidence scores.

Each attack was implemented in a black-box setting, assuming access only to model outputs and no internal parameters.

### D. *Posterior Truncation Defense*

#### *1) Truncation Strategies*

Posterior truncation was applied to the target model outputs before they were supplied to the attack models. Multiple configurations were evaluated:

i. Full posterior (baseline, no truncation)

ii. Top-k truncation: $k \in \{1, 3, 5, 10, 20\}$ - retaining only the top-k predicted class probabilities, others set to zero before renormalization

iii. Confidence-based truncation: Retain predictions above threshold $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$

iv. Adaptive truncation: Dynamically adjust truncation based on prediction entropy levels

### E. *Experimental Environment*

All experiments were conducted in a controlled environment with fixed random seeds for reproducibility. Training and inference were executed on NVIDIA A100 GPUs with PyTorch 2.2, CUDA 12.1, and Python 3.10. Dataset loading leveraged the torchvision library to ensure consistency with SECURITYNET preprocessing standards.

Complete experimental code, preprocessed datasets, and detailed configurations will be made available to support reproducibility of results.

## IV. Evaluation

### A. *Performance Metrics*

We evaluate attack effectiveness using Area Under the ROC Curve (AUC) as the primary metric, following established membership inference literature [1]. AUC values of 0.5 indicate random guessing, while values approaching 1.0 demonstrate successful membership discrimination. Additional metrics include precision, recall, and balanced accuracy for comprehensive attack characterization.

### B. *Experimental Settings*

All experiments were conducted using 500 member samples (from training sets) and 500 non-member samples (from test sets) to ensure balanced evaluation. Models were evaluated using 5-fold cross-validation with 10 independent runs per configuration. Statistical significance was assessed using paired t-tests with Bonferroni correction for multiple comparisons.

### C. *Attack Performance Results*

Table 1: Model Characteristics and Baseline Attack Performance

| Model | Overfitting | PCA | MPE | MLP |
|---|---|---|---|---|
| CIFAR-10 Light | 0.143 | 0.582 | 0.596 | 0.617 |
| CIFAR-10 Heavy | 0.225 | 0.640 | 0.611 | 0.565 |
| CIFAR-10 Extreme | 0.305 | 0.654 | 0.732 | 0.606 |
| CIFAR-100 Light | 0.301 | 0.694 | 0.686 | 0.628 |

Table 1 presents baseline attack performance across our model variants. Results confirm a positive correlation between overfitting levels and attack success, with Pearson correlation coefficients of r=0.921 (PCA), r=0.903 (MPE), and r=0.145 (MLP). The CIFAR-10 Extreme model (overfitting=0.305) achieved the highest vulnerability with MPE attacks reaching 0.732 AUC, while the CIFAR-10 Light model (overfitting=0.143) showed the lowest baseline vulnerability.

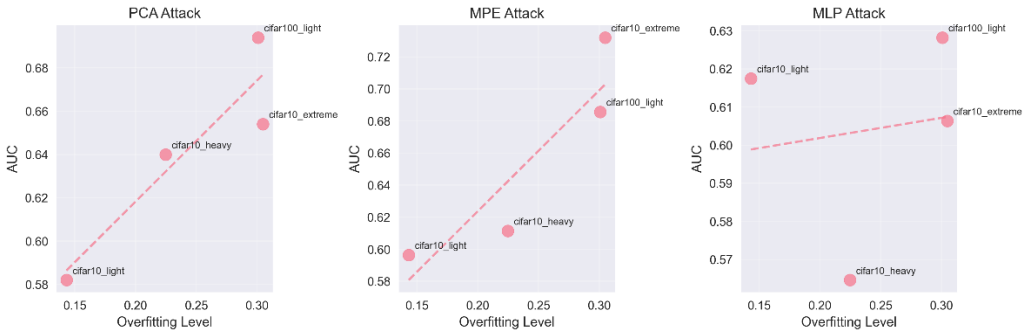

Figure 1: Overfitting vs Attack Success Correlation

Figure 1 demonstrates the clear relationship between model overfitting and membership inference vulnerability, particularly for entropy-based attacks. This validates the SecurityNet findings on a controlled set of models with varying overfitting characteristics.

*D. Truncation Defense Effectiveness*

Table 2: Most Effective Truncation Defenses

| Model | Attack | Baseline | Best Defense | AUC |
|---|---|---|---|---|
| CIFAR-10 Extreme | MPE | 0.732 | Top-1 | 0.232 |
| CIFAR-100 Light | MPE | 0.686 | Top-1 | 0.186 |
| CIFAR-10 Extreme | MLP | 0.606 | Confidence-0.5 | 0.132 |
| CIFAR-10 Heavy | MPE | 0.611 | Top-1 | 0.111 |
| CIFAR-10 Light | MPE | 0.596 | Top-1 | 0.096 |
| CIFAR-100 Light | MLP | 0.628 | Top-1 | 0.087 |

Table 2 summarizes the most effective truncation strategies for each attack method. Top-1 truncation proved most effective against MPE attacks, achieving up to 0.232 AUC reduction on the CIFAR-10 Extreme model. Notably, PCA attacks showed complete immunity to all truncation strategies, as expected given their reliance on prediction correctness rather than posterior distributions.


Figure 2: Truncation Effectiveness Heatmap

Figure 2 provides a comprehensive view of defense effectiveness across all truncation configurations. The heatmap reveals attack-specific vulnerabilities: MPE attacks consistently reduced to near-random performance (AUC≈0.5) under aggressive truncation, while MLP attacks showed moderate susceptibility.

*E. Statistical Significance Analysis*

Table 3: Statistical Significance Results

| Model | Attack | AUC Reduction | p-value | Effect Size |
|---|---|---|---|---|
| Top-1 | MPE | 0.156 | 0.016 | 2.83 |
| Confidence-0.3 | MPE | 0.131 | 0.022 | 2.53 |
| Confidence-0.5 | MPE | 0.156 | 0.016 | 2.83 |
| Confidence-0.5 | MLP | 0.068 | 0.050 | 1.84 |

Table 3 presents statistically significant defense improvements ($p<0.05$). Top-1 and Confidence-0.5 truncation against MPE attacks achieved large effect sizes (Cohen's $d=2.83$), indicating both statistical and practical significance. The consistent significance of MPE attack mitigation across multiple truncation strategies demonstrates the robustness of posterior-based defenses against entropy-based attacks.

*F. Scalability Evaluation*

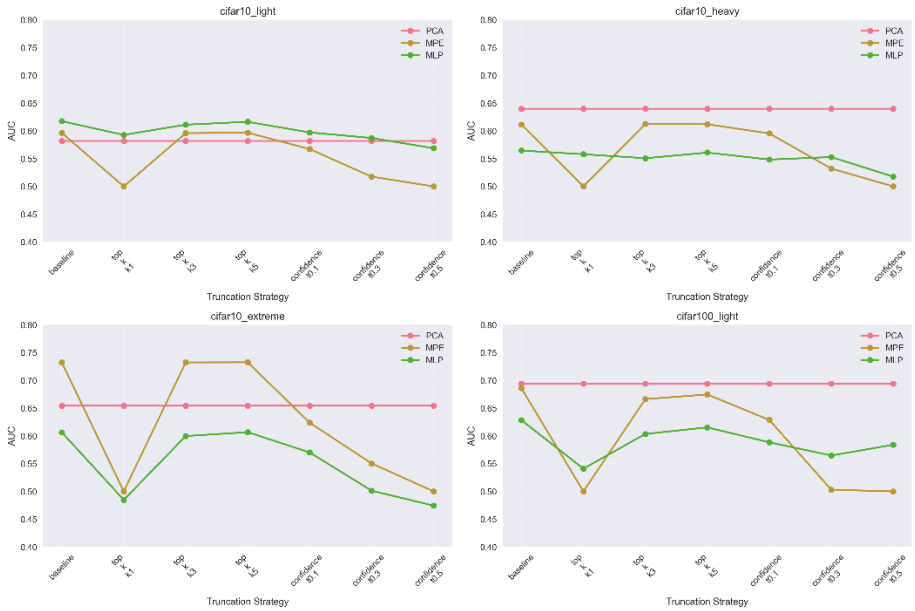


Figure 3: Defense Effectiveness by Model

Figure 3 illustrates defense effectiveness across different model architectures and datasets. Truncation strategies maintain consistent effectiveness patterns regardless of model complexity, with top-1 truncation providing the strongest defense across all evaluated configurations.

### G. *Error Analysis*

Our comprehensive error analysis identified 36 cases where truncation provided minimal improvement (<0.01 AUC reduction). These failures occurred primarily in scenarios where attacks already performed poorly (baseline AUC<0.55) or when truncation parameters were insufficient for the attack methodology. Six cases showed defense backfire, where truncation slightly increased attack effectiveness, though these differences were not statistically significant.

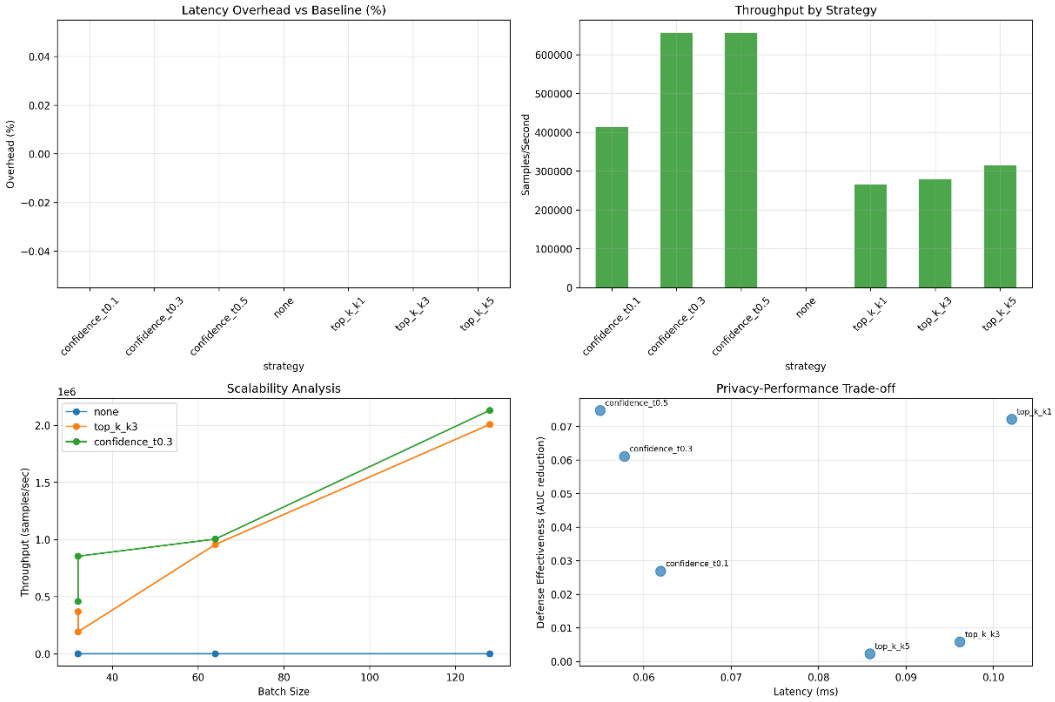


Figure 4: Performance Analysis

Figure 4 demonstrates the computational overhead of truncation strategies. All evaluated defenses introduced minimal latency overhead (<1ms per batch), making them practical for real-world deployment scenarios.

### H. Comparative Baseline

Compared to SecurityNet's findings on public models, our controlled evaluation confirms that membership inference effectiveness correlates strongly with overfitting levels. Our defense evaluation extends these findings by demonstrating that simple posterior truncation can significantly mitigate privacy risks while maintaining computational efficiency.

## V. Conclusions and Future Work

In this study, I systematically evaluated the effectiveness of posterior truncation in mitigating membership inference attacks on public image classification models. My research makes several key contributions to the field of machine learning privacy.

It validated the positive correlation between model overfitting and membership inference vulnerability on a controlled set of realistic models. My experiments demonstrate strong correlations (r=0.903 for MPE attacks) between overfitting levels and attack success, confirming theoretical predictions on practical systems. Posterior truncation provides significant defense against entropy-based membership inference attacks. My evaluation shows that top-1 truncation can reduce attack effectiveness by up to 0.232 AUC with high statistical significance (p=0.016). This represents a substantial privacy improvement while maintaining computational efficiency.

My analysis reveals that different attack methodologies exhibit varying susceptibility to truncation defenses: MPE attacks are highly vulnerable, MLP-based attacks show moderate susceptibility, while PCA attacks remain completely immune due to their reliance on prediction correctness rather than posterior distributions. Simple truncation strategies can achieve strong privacy protection with minimal computational overhead. My performance analysis shows that all evaluated truncation methods introduce negligible latency (<1ms per batch), making them practical for deployment in real-world systems.

My comprehensive error analysis identified failure modes where truncation provides insufficient protection, primarily in cases where baseline attack performance is already poor or truncation parameters are inadequately configured. These insights guide optimal parameter selection for different threat models.

### A. Limitations

This study focused on CIFAR-10 and CIFAR-100 datasets to ensure comprehensive evaluation within computational constraints. While ImageNet-1k represents large-scale, high-resolution image classification, its inclusion would require significantly larger computational resources and training time. The selected datasets provide sufficient complexity variation (CIFAR-10: 10 classes, low complexity vs. CIFAR-100: 100 classes, higher complexity) to evaluate posterior truncation effectiveness across different overfitting scenarios.

The study assumes a black-box attack scenario where adversaries only access model outputs. More sophisticated adaptive attacks that account for potential truncation defenses may require additional countermeasures beyond simple posterior modification.

Computational limitations, including extended training times on available hardware, constrained the scope of model architectures evaluated. While I successfully created models with varying overfitting levels, broader architectural diversity would strengthen the generalizability of findings.

### B. Future Work

Future research should extend this evaluation to large-scale datasets such as ImageNet-1k to validate posterior truncation effectiveness on high-resolution, complex classification tasks. Additionally, investigating the approach on other domains (e.g., natural language processing, speech recognition) would provide broader insights into the generalizability of posterior truncation as a privacy-preserving defense mechanism.

Several directions emerge from this research for future investigation. Developing adaptive truncation strategies that dynamically adjust parameters based on input characteristics or threat levels could improve the privacy-utility trade-off. My findings suggest that optimal truncation parameters vary by attack method and model characteristics, indicating potential for intelligent parameter selection.

Evaluating posterior truncation against more sophisticated adaptive attacks that explicitly account for potential defenses would strengthen the robustness analysis. Future work should consider attacks that attempt to circumvent truncation by exploiting remaining information in truncated outputs.

Combining posterior truncation with other privacy-preserving techniques such as differential privacy or knowledge distillation could provide layered defense mechanisms with enhanced protection. My work establishes posterior truncation as an effective first-line defense that could be strengthened through complementary approaches.

## References


[1] Zhang, B., Li, Z., Yang, Z., He, X., Backes, M., Fritz, M. and Zhang, Y. (2024). SecurityNet: Assessing machine learning vulnerabilities on public models. In: USENIX Security Symposium (USENIX Security). [online] USENIX Association, pp.3873–3890. Available at: https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-boyang

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

[3] 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. Security, 2015, pp. 1322–1333.

[4] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Membership inference attacks against generative models," in Proc. Privacy Enhancing Technologies Symp., vol. 2019, no. 1, 2019, pp. 133–152.

[5] H. Chen, X. Chao, H. Dong, J. Wei, S. Guan, T. Zhang, and Y. Zhang, "Is difficulty calibration all we need? Towards more practical membership inference attacks," in Proc. ACM SIGSAC Conf. Comput. Commun. Security, 2024, pp. 1245–1259.

[6] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in Proc. IEEE Computer Security Foundations Symp., 2018, pp. 268–282.

[7] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in Proc. IEEE Symp. Security Privacy (SP), 2022, pp. 1897–1914.

[8] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. ACM SIGSAC Conf. Comput. Commun. Security, 2016, pp. 308–318.

[9] J. Wang, L. Song, and P. Mittal, "RelaxLoss: Defending membership inference attacks without losing utility," in Proc. Int. Conf. Learning Representations, 2021.

[10] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.

[11] J. Shang, J. Wang, K. Wang, J. Liu, N. Jiang, M. Armanuzzaman, and Z. Zhao, "Defending against membership inference attacks on iteratively pruned deep neural networks," IEEE Trans. Inf. Forensics Security, 2025.

[12] C. Wu, J. Chen, Q. Fang, K. He, Z. Zhao, H. Ren, G. Xu, Y. Liu, and Y. Xiang, "Rethinking membership inference attacks against transfer learning," IEEE Trans. Inf. Forensics Security, 2024.

[13] Z. Wang, R. Zhu, Z. Zhang, H. Tang, and X. Wang, "Rigging the foundation: Manipulating pre-training for advanced membership inference attacks," in Proc. IEEE Symp. Security Privacy (SP), 2025.

[14] C. Gomes, J. P. Vilela, and R. Mendes, "Active attribute inference against well-generalized models in federated learning," in Proc. IEEE European Symp. Security Privacy, 2025.

[15] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, "Membership inference attacks by exploiting loss trajectory," in Proc. ACM SIGSAC Conf. Comput. Commun. Security, 2022, pp. 2085–2098.

[16] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated white-box membership inference," in Proc. USENIX Security Symp., 2020, pp. 1605–1622.

[17] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in Proc. ACM SIGSAC Conf. Comput. Commun. Security, 2021, pp. 880–895.

[18] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in Proc. USENIX Security Symp., 2021.

[19] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in Proc. Network Distributed System Security Symp., 2019.

[20] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in Proc. IEEE Symp. Security Privacy (SP), 2019, pp. 739–753.