Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082713

ARTICLE

HEbdMIA: Lightweight Logit Encryption for Membership Inference Defense

Akash Shah[1], Mudasir Ahmad Wani[2,*], Ravi Prakash Chaturvedi[3], Shri Kant[3], Nidhi Sindhwani[1], Kashish Ara Shakil[4] and Sulieman Alshuhri[2]

[1]Amity Institute of Information Technology, Amity University, Noida, Uttar Pradesh, India
[2]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
[3]Center for Cyber Security and Cryptology, School of Computing Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India
[4]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

*Corresponding Author: Mudasir Ahmad Wani. Email: mawani@imamu.edu.sa

Received: 21 March 2026; Accepted: 13 May 2026; Published: 23 July 2026

ABSTRACT: Membership Inference Attacks (MIAs) pose a significant privacy risk in machine learning by enabling adversaries to infer whether specific data samples were used during training, particularly in sensitive domains such as social media and mental health analytics. To address this challenge, this paper proposes HEbdMIA, a lightweight homomorphic encryption-based defense that operates at the post-inference stage by encrypting model output logits without requiring retraining or architectural modifications. The proposed approach preserves the relative ordering of predictions while obscuring confidence patterns exploited by MIAs. Experimental evaluation on DepInferAttack and BotInferAttack demonstrates that HEbdMIA achieves a reduction in MIA success rates of 31.0% and 27.3%, respectively, with an associated accuracy decrease of 29.3% and 26.4%, reflecting a controlled privacy and utility trade off. Additional analysis using precision, recall, F1-score, and ROC-AUC confirms a substantial decline in adversarial inference capability. These findings indicate that HEbdMIA provides an effective, scalable, and deployment-friendly solution for enhancing privacy in real-world machine learning systems.

KEYWORDS: Membership inference attack; homomorphic encryption; machine learning; encrypted data; logit encryption; model security

1 Introduction

Machine learning (ML) models have become integral to various privacy-sensitive domains such as healthcare, online social networks (OSNs), and personalized recommendation systems [1,2]. Despite their success, these models are vulnerable to Membership Inference Attacks (MIAs) a class of privacy attacks that allow an adversary to infer whether a given data point was used in the model's training phase [3]. Such attacks are possible because many models exhibit distinguishable behaviors (e.g., higher confidence) for training data compared to unseen data. Since Shokri et al.'s foundational work [3], MIAs have been shown to be effective across a wide range of architectures, including convolutional neural networks (CNNs), recurrent models, and transformer-based models [4–7].

1.1 Membership Inference Attacks

Membership inference attacks (MIAs) represent a major privacy concern in modern machine learning, particularly when models are deployed in domains involving sensitive data such as medical records, user behavior analytics, or mental health prediction. First introduced by Shokri et al. [3], MIAs allow adversaries to determine whether a specific data point was included in the training set of a machine learning model by analyzing differences in model output behavior, such as prediction confidence, margin, or entropy between member and non-member samples.

The classical MIA framework is composed of three key components: the target model, the shadow model, and the attack model. The target model is the deployed machine learning system under attack [8]. It is trained on sensitive data and exposed via a query interface (e.g., an API), typically providing softmax probabilities or logits for any input query. This is the primary point of vulnerability [9,10].

To simulate the target model's behavior, the attacker constructs one or more shadow models. These are trained on a separate dataset drawn from the same or similar distribution. By controlling the shadow training process, the adversary knows which samples are members and which are not, thereby creating labeled data for attack model training. Recent research [11] has even demonstrated using shadow-free approaches under label-only conditions.

The attack model is a binary classifier trained using the outputs of the shadow models. It learns to differentiate between member and non-member samples based on statistical signals in the output, such as class probabilities, prediction entropy, or gradients [12].

1.2 Motivation

MIAs pose serious privacy risks in sensitive domains such as mental health analysis and social network monitoring, where identifying training data membership can expose personal or critical information. These attacks exploit confidence and statistical patterns in model outputs to distinguish training samples. Existing defenses, including differential privacy, adversarial regularization, and output perturbation, provide partial protection but often suffer from utility loss, retraining requirements, and high computational overhead [13,14], while lacking flexibility for cross-domain and black-box deployment. As MIAs evolve into more adaptive forms [15,16], there is a growing need for lightweight, post-inference defenses that effectively obscure such leakage. HEbdMIA addresses this gap by applying homomorphic encryption to output logits, preserving model architecture while providing efficient and practical privacy protection without retraining.

1.3 HEbdMIA: Homomorphic Encryption-Based Defense against Membership Inference Attacks

To address this growing concern, this paper proposes a novel framework called HEbdMIA Homomorphic Encryption-based Defense Against Membership Inference Attacks. Unlike traditional defenses that rely on differential privacy [13], adversarial training [17], or noise injection [18]. HEbdMIA introduces a homomorphic encryption [19] layer applied to the model's output logits during inference. This transformation preserves prediction accuracy while disrupting the statistical signals that adversaries exploit during attack reconstruction.

HEbdMIA is designed to be model-agnostic and deployable without retraining. It is evaluated against two real-world MIA models:

- DepInferAttack [1]: Perform MIA on machine learning models that are trained on Reddit-based mental health data.
- BotInferAttack [2]: Perform MIA on machine learning models that are trained to detect bots on OSN datasets.

1.4 Logits

In the context of deep learning models, the term logits refer to the raw, unnormalized outputs of the final layer of a neural network, produced before the application of a normalization function such as softmax. Formally, given an input sample x, the model computes a vector of real-valued scores $z = (z_1, z_2, \ldots, z_k)$, where k is the number of classes. These values represent the model's confidence in each class prior to conversion into probabilities:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}, i \in \{1, \ldots, k\} \tag{1}$$

Membership inference attacks often exploit the distribution of logits to distinguish between member and non-member samples. This makes logits a critical focal point for privacy-preserving defenses, motivating our use of homomorphic encryption to securely transform them in HEbdMIA.

1.5 Contributions

This paper makes the following key contributions:

- Proposed HEbdMIA a novel homomorphic encryption based defense at the logit level, effectively mitigating membership inference leakage without requiring model retraining or architectural modifications.
- Establishes the robustness and generalizability of the proposed defense through evaluation on two distinct and privacy-sensitive application domains: DepInferAttack [1] for depression detection in social media text and BotInferAttack [2] for bot detection in online social networks.
- The experimentation results show a reduction of around 30% in membership inference attack success, consistently observed across confidence, entropy, and logit-based attack models.
- Provides a detailed computational efficiency analysis, showing that post-inference logit encryption incurs minimal overhead, supporting deployment in real-time and resource-constrained environments.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the proposed HEbdMIA framework, Section 4 describes the experimental setup, Section 5 discusses the results, and Section 6 concludes the study with future directions.

2 Literature Review

Recent advancements in membership inference attacks and privacy-preserving machine learning have improved both attack strategies and defense mechanisms. Existing defenses can be broadly categorized into confidence-regularization, output-space transformation, and encryption-based approaches, each addressing membership leakage from different perspectives.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the proposed HEbdMIA framework, Section 4 describes the experimental setup, Section 5 discusses the results, and Section 6 concludes the study with future directions.

2.1 Confidence-Regularization Methods

Confidence-regularization techniques aim to curb membership leakage by constraining the model's output confidence during training. HAMP (High Accuracy and Membership Privacy) proposed by Chen and Pattabiraman (2023) introduces an entropy-based loss that encourages the model to maintain higher uncertainty across classes, reducing MIA success while retaining high utility [20]. Concurrently, Li et al. (2023) introduced MIST (Membership-Invariant Subspace Training), which reshapes the feature space so

that representations of different instances are more uniformly distributed, specifically targeting vulnerable samples to mitigate membership attacks without affecting model architecture [21]. Earlier work, such as the DP-SGD framework by Abadi et al. (2016), protects against MIAs by injecting noise into gradient updates and clipping their norms; while effective, this approach often entails significant performance trade-offs [13]. Nasr et al. (2018) introduced an adversarial regularization approach where a membership inference attacker is incorporated during training, allowing the model to reduce membership leakage while maintaining prediction accuracy [17]. Similarly, Zhang et al. (2024) proposed Repair-Membership Learning, which adjusts output distributions using adaptive soft labels to minimize confidence gaps exploited by membership inference attacks while preserving predictive performance [22]. Recent studies, HAMP [20], MIST [21], and Repair Membership Learning [22], have improved confidence regularization techniques by reducing generalization gaps and controlling prediction certainty. However, these approaches require retraining and are not applicable in post-deployment scenarios.

2.2 Output-Space Transformation Techniques

While training-time defenses provide strong theoretical guarantees, they often come at the cost of training flexibility. To this end, researchers have explored post-inference or output-level defenses. One of the most well-known methods is MemGuard, proposed by Jia et al. (2019), which perturb model logits using adversarial noise crafted to mislead attack models [18]. MemGuard is particularly effective in black-box settings and does not require model retraining, making it a practical defense option. However, later evaluations revealed that adaptive attackers who model the perturbation process can reduce its effectiveness. To overcome these limitations, some recent studies have attempted to increase generalization by modifying prediction behaviors. For example, AdaMixup [23], proposed in 2025, uses adaptive interpolation during training to produce smoother decision boundaries and reduce overfitting. Although it shows improved resistance to MIAs, AdaMixup [23] still fundamentally depends on retraining and does not address inference-time leakage directly. Yadav and Mane (2025) proposed a metric mapping method that modifies entropy and probability distributions to reduce membership inference attacks while maintaining accuracy [24]. Yang et al. (2025) introduced a diffusion-based preprocessing technique that removes membership-sensitive signals from outputs with minimal impact on classification performance [25]. More recent works, such as metric mapping [24] and diffusion-based output obfuscation [25], attempt to hide membership signals through output transformations. While these approaches improve robustness against certain attacks, they may introduce instability or distort prediction confidence.

2.3 Encryption-Based Approaches in Machine Learning Privacy

Homomorphic encryption (HE) has emerged as a promising approach for protecting privacy in machine learning, particularly in contexts involving federated learning and secure inference. Early efforts, such as those by Gentry [26], laid the foundation for fully homomorphic encryption (FHE), although initial implementations were computationally impractical. More recently, frameworks such as HE-MAN [27] have demonstrated the feasibility of running encrypted inference on ONNX models, but these systems focus primarily on preserving input and model confidentiality rather than defending against inference-based privacy attacks like MIAs. ReActHE [28] extended the application of HE to biomedical predictions, creating deep neural network architectures tailored to operate efficiently under encryption. Byun et al. (2024) proposed privacy-preserving deep learning frameworks that integrate secure computation with regularization techniques to protect sensitive inputs and reduce risks from model extraction and membership inference attacks while maintaining accuracy [29]. Similarly, Wagh et al. (2019) introduced SecureNN, which

uses secret sharing and secure computation protocols to enable collaborative neural network inference without exposing sensitive data [30].

Recent advancements in homomorphic encryption-based machine learning have focused on improving the efficiency of encrypted inference and secure computation frameworks. However, these approaches primarily protect input data and model parameters rather than addressing information leakage from model outputs, leaving a critical gap in defending against membership inference attacks.

Table 1 shows a comparative summary of existing membership inference attack defense mechanisms based on confidence-based data masking, output perturbation, and encryption-based techniques.

Table 1: Comparative summary of membership inference attack defenses.

Approach Type	Method	Defense Strategy	Retraining Required	Post-Inference Applicable	Limitation
Confidence Based Data	HAMP [20]	High-entropy regularization	Yes	No	Requires access to training pipeline
	MIST [21]	Membership-invariant subspace projection	Yes	No	Not applicable post-deployment
	Adversarial Regularization [17]	Joint adversarial training with a simulated MIA attacker	Yes	No	Computationally expensive and requires adversarial training setup
	Repair Membership Learning [22]	Soft-label repair learning to reduce loss gap between members and non-members	Yes	No	Requires retraining and modification of training labels
Output Perturbation	MemGuard [18]	Adversarial perturbation of logits	No	Yes	Vulnerable to adaptive attacks
	AdaMixup [23]	Adaptive mixup with dynamic weights & labels	Yes	No	Requires full model retraining
	Metric Mapping Defense [24]	Maps output confidence scores to alternative metric space to hide membership signals	No	Yes	May distort confidence calibration of predictions
	Diffusion-Based Output Obfuscation [25]	Adds controlled stochastic noise through diffusion-style transformations to obscure membership patterns	No	Yes	Computational overhead and may affect prediction stability
	HE-MAN [27]	Homomorphic encryption of model inputs	Yes	No	Does not protect inference outputs

(Continued)

Table 1 (continued)

Approach Type	Method	Defense Strategy	Retraining Required	Post-Inference Applicable	Limitation
Encryption Based Technique	ReActHE [28]	Homomorphic Encryption optimized model architecture	Yes	No	Architecture-specific
	Secure Privacy-Preserving Inference Framework [29]	Uses secure computation protocols and cryptographic techniques to perform inference on protected data while preventing information leakage	Yes	No	High computational and communication overhead
	SecureNN [30]	Secure multiparty computation for privacy-preserving inference	Yes	No	High computational and communication overhead

2.4 Limitations of Existing Approaches and Rationale for Proposed Work

Existing defenses against Membership Inference Attacks (MIAs) primarily rely on model retraining (e.g., HAMP [20], MIST [21]) or heuristic output perturbations (e.g., MemGuard [18]), often leading to utility degradation or reduced robustness against adaptive adversaries. Similarly, homomorphic encryption-based approaches such as HE-MAN [27] and ReActHE [28] focus on protecting inputs or model computations, while neglecting leakage from model outputs.

This reveals a critical gap in developing lightweight, post-inference defenses that can mitigate membership leakage without modifying the model architecture. Existing methods are also often domain-specific and lack general applicability across different data modalities. However, like most HE applications in ML, these systems are concerned with the secure transmission and processing of data, not with the leakage of information from model outputs post-inference. The lack of application of HE specifically for post-inference MIA protection remains a significant research gap.

To address these limitations, we propose HEbdMIA, a post-inference defense that applies homomorphic encryption directly to output logits. Unlike perturbation-based methods such as MemGuard, which rely on adversarial noise, HEbdMIA provides cryptographic protection that preserves prediction ordering while preventing leakage of confidence patterns. Our proposed approach, HEbdMIA, directly addresses this limitation by applying homomorphic encryption at the logit level post-inference, thereby mitigating leakage without retraining. This design avoids retraining, introduces minimal overhead, and offers a scalable, model-agnostic solution suitable for real-world deployment across diverse applications.

3 Methodology

The following section presents the experimental setup used to evaluate the effectiveness of the proposed HEbdMIA framework under different attack scenarios. The primary objective of HEbdMIA is to defend against Membership Inference Attacks (MIAs) by encrypting the output logits of a pre-trained classifier during the post-inference phase, thereby eliminating the need to alter the model architecture or retrain

the model. HEbdMIA is designed to be compatible with any classification task and preserves predictive performance while significantly reducing the leakage of membership information.

3.1 HEbdMIA Architecture

The HEbdMIA pipeline is composed of three core components: a trained target model, a post-inference encryption module, and a secure prediction mechanism. For any input sample x, the model $f(.)$ produces an output logit vector $\boldsymbol{z} = f(x) \in \mathbf{R}^k$, where k is the number of output classes. These logits $\boldsymbol{z} = [z_1, z_2, ..., z_k]$ represent the unnormalized class scores generated by the final layer of the model.

To obscure the statistical structure of the logits from an adversary, each element of the logit vector is encrypted using a partially homomorphic encryption scheme. In this work, we employ the Paillier encryption algorithm, which supports additive homomorphism and allows computations over encrypted values without revealing the plaintext. The encrypted logits are represented as:

$$\mathbf{z}' = Enc(z) = [Enc(z_1), Enc(z_2), \ldots, Enc(z_k)] \tag{2}$$

This ciphertext vector $\boldsymbol{z}'$ is not directly accessible to external querying agents. Instead, it is either retained within a secure execution environment or returned to a trusted client that holds the decryption key for final prediction and interpretation.

After defining the encrypted logit vector, the encryption process in the Paillier cryptosystem can be formally expressed. For any plaintext logit value z_i, the encryption function is defined as:

$$Enc(z_i) = g^{z_i} \cdot r^n \bmod n^2 \tag{3}$$

where g is the generator of the public key, nis the RSA modulus, and r is a randomly selected number such that $r \in \mathbb{Z}_n^*$. The randomness r ensures that even identical logits produce different ciphertext values, thereby preventing statistical leakage.

In this work, we employ the Paillier cryptosystem [31], a partially homomorphic encryption scheme that supports additive operations over encrypted data. Unlike fully homomorphic encryption schemes such as BFV or CKKS, which enable arbitrary computations over ciphertexts but incur high computational overhead, Paillier offers a more efficient alternative for scenarios where only limited operations are required. Since HEbdMIA operates exclusively on low-dimensional output logits and does not require complex encrypted computations, the use of Paillier ensures minimal latency and practical deployability while maintaining strong semantic security.

3.2 Post-Inference Encryption and Prediction

HEbdMIA performs encryption on the model's logits after inference, in contrast to conventional schemes (HE-MAN [27]), which encrypt the input data or the entire model computation. This design choice enables model outputs to be protected without requiring retraining or altering the learning pipeline. Once the encrypted logits $\boldsymbol{z}'$ are generated, class prediction is achieved by decrypting the logits and selecting the index with the highest score.

The predicted class label $\hat{y}$ is computed as:

$$\hat{y} = \arg \sum_{i \in \{1,\ldots,k\}} Dec(\boldsymbol{z}_i^{'}) \tag{4}$$

In this expression, $Dec(z_i^{'})$ denotes the decryption function associated with the Paillier cryptosystem. Decryption is performed in a secure environment, ensuring that only the final prediction $\hat{y}$ is exposed. This prevents adversaries from accessing logits or confidence-based information used in membership inference attacks.

Finally, the trusted client holding the private key performs decryption as:

$$Dec(z_i^{'}) = \frac{L(c_i^{\lambda} \bmod n^2)}{L(g^{\lambda} \bmod n^2)} \bmod n \tag{5}$$

where λ represents the private key parameter.

Through this mechanism, the HEbdMIA framework transforms the original logit vector into a protected ciphertext space before exposure to external queries, thereby preventing attackers from exploiting confidence distributions typically used in membership inference attacks.

3.3 Threat Model

The proposed HEbdMIA framework operates under a black-box threat model, where the attacker can query the deployed model with arbitrary inputs and observe its predictions. The attacker may possess auxiliary data sampled from the same distribution as the training set and aims to determine whether a specific sample was included in the training data of the target model. This assumption is consistent with the attack scenarios used in DepInferAttack [1] and BotInferAttack [2], where the adversary leverages statistical differences in the model's output logits to perform inference.

We assume that the attacker does not have access to the encryption key or any internal model parameters. The homomorphic encryption scheme used in HEbdMIA guarantees semantic security, meaning that ciphertexts do not reveal information about the plaintext values. Consequently, the adversary cannot distinguish between members and non-members based on the encrypted logits, even if they mount adaptive attacks using confidence thresholding or statistical classifiers.

In addition to the standard black-box setting, we consider a relaxed scenario where the adversary may have access to partial outputs, such as confidence scores. Even in such cases, HEbdMIA mitigates membership inference risks by encrypting the output logits, thereby preventing the exposure of statistical patterns required for successful inference.

3.4 HEbdMIA Workflow

The proposed HEbdMIA framework operates under a black-box threat model, where the adversary queries the model and observes output logits or probability scores, aiming to infer membership using auxiliary data, as in DepInferAttack [1] and BotInferAttack [2]. Attacks such as confidence thresholding, entropy-based inference, and logit-based classification exploit statistical differences in model outputs, particularly higher confidence for training samples. The adversary is assumed to be semi-honest, with no access to model parameters, training data, or encryption keys. To mitigate these risks, HEbdMIA applies homomorphic encryption to output logits before exposure. Due to the semantic security of the encryption scheme, ciphertexts reveal no exploitable information, preventing inference of confidence patterns even under adaptive attacks. The overall workflow is illustrated in Fig. 1.

To evaluate the effectiveness of HEbdMIA, we employ three widely used membership inference strategies confidence thresholding, entropy-based inference, and logit-based classifiers which represent diverse attack perspectives in MIA research. Confidence thresholding exploits higher prediction confidence for training samples [32], while entropy-based inference leverages lower uncertainty (entropy) associated

with such samples [14]. Logit-based classifiers represent a stronger attack by learning discriminative patterns directly from raw logits [33]. HEbdMIA mitigates these attacks by encrypting logits, thereby obscuring confidence, entropy, and fine-grained output patterns. This reduces the separability between member and non-member samples while preserving predictive utility, enabling a comprehensive and robust evaluation across diverse attack scenarios.

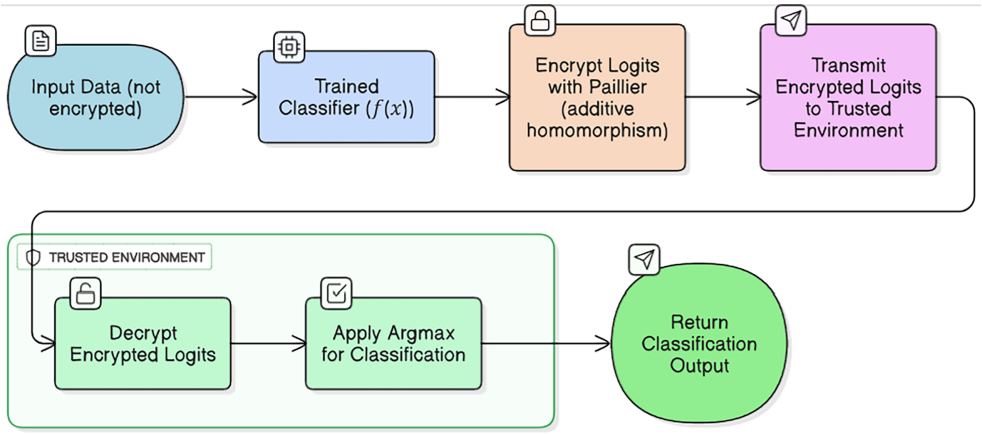

Figure 1: Workflow of the HEbdMIA framework.

To preserve confidentiality of model outputs during transmission and aggregation, we employ the Paillier cryptosystem, which provides additive homomorphic encryption. Let the model produce a real-valued logit vector $L = [l_1, \ldots, l_k] \in \mathbf{R}^k$. Prior to encryption, logits are scaled and quantized into integer representations to satisfy the integer-domain requirement of Paillier encryption. Using a single key generation setup, a public–private key pair (pk, sk) is generated, where the public key is shared for encryption and the private key is securely retained for decryption. Each logit is then encrypted independently, enabling secure aggregation directly in the ciphertext domain. The detailed encryption procedure is shown in Fig. 2.

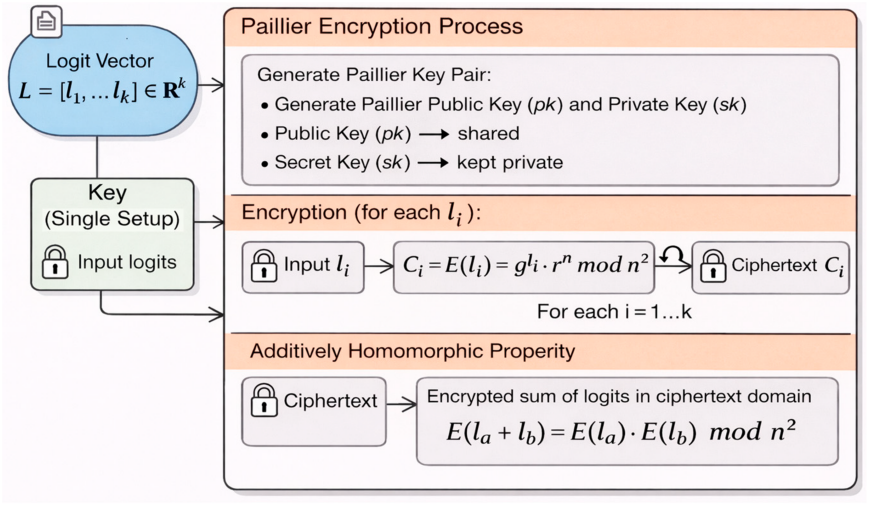

Figure 2: Detailed logic of paillier encryption for logits.

Algorithm 1 presents the detailed operational workflow of the proposed HEbdMIA framework. It outlines how the defense is applied in the post-inference stage by securely transforming the output logits of a trained target model using homomorphic encryption. This algorithm explicitly captures the sequence of steps involved in protecting inference outputs while preserving the original prediction process, thereby mitigating membership inference attacks without requiring model retraining or architectural modifications.

Algorithm 1: HEbdMIA–post-inference logit encryption defense

Input: M: Trained target model and x: input sample
Output: Predicted class label y' and encrypted logits z'
procedure HEbdMIA (M, x)
$z \leftarrow M(x)$//Forward inference to obtain logits
$z_{norm} \leftarrow$ Normalize (z)//Normalize logits for stability
for each logit $z_i \in z_{norm}$ **do**
$z_i' \leftarrow$ HE_Encrypt (z_i)//Apply homomorphic encryption
end for
$z' \leftarrow \{z_i'\}$//Encrypted Logit vector
$z_{decrypt} \leftarrow$ HE_Decrypt (z')//Decrypt logits for prediction
$y' \leftarrow$ argmx$\{z_{decrypt}\}$//Final class decision
return y', z'
end procedure

3.5 Efficiency and Deployment Considerations

HEbdMIA is designed for practical deployment in privacy-sensitive applications such as social network analysis and mental health classification. Since encryption is applied only to low-dimensional logits, the computational and memory overhead remains minimal. Unlike retraining based or full model encryption approaches, HEbdMIA requires no modification to the underlying model and can be seamlessly integrated into existing systems. This lightweight, plug-and-play design makes it suitable for real-world deployment under privacy regulations such as GDPR [34] and HIPAA.

3.6 Security Analysis

HEbdMIA mitigates key vulnerabilities exploited by Membership Inference Attacks by encrypting output logits at the post-inference stage, thereby preventing access to confidence scores, entropy, and logit margins used for inference. The semantic security of the Paillier scheme ensures that no meaningful information is revealed even under repeated queries, making the approach robust against both threshold-based and shadow model attacks. By suppressing statistical signals in model outputs, HEbdMIA effectively prevents discrimination between member and non-member samples, as validated by consistent reductions in attack success rate, membership advantage, and ROC-AUC across DepInferAttack [1] and BotInferAttack [2].

4 Experimental Setup

This section presents the experimental framework used to evaluate the proposed HEbdMIA defense against membership inference attacks. We revisit two previously established attack scenarios DepInferAttack [1] and BotInferAttack [2] where prior work demonstrated that adversaries could exploit model outputs to infer training data membership. Here, we adopt these same scenarios to rigorously test HEbdMIA's effectiveness in mitigating those vulnerabilities, without compromising model performance. The section details the datasets used, preprocessing techniques, target model configurations, and the setup for executing

membership inference attacks. All experiments were conducted on a system equipped with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA GPU.

4.1 Dataset Description

4.1.1 Dataset 1: DepInferAttack

The second scenario is based on the DepInferAttack framework [1], where we previously demonstrated membership inference vulnerabilities in deep learning models trained for mental health classification on social media. The dataset consists of 7731 Reddit posts, collected using the official Reddit API. Posts were extracted from mental health-related subreddits and general forums and then labeled using a keyword-based approach. Keywords indicative of depressive states (e.g., sad, worthless, hopeless) were used to identify depressed samples, while neutral or positive terms (e.g., cheerful, productive) labeled normal samples.

The final dataset includes 3831 depressed and 3900 normal samples. In our prior attack study, models trained on this dataset were shown to be highly susceptible to MIA, largely due to distinct lexical and semantic patterns that were retained in the model's logits. Fig. 3 presents a bar chart showing the number of instances of both normal and depressed text used in the DepInferAttack Model.

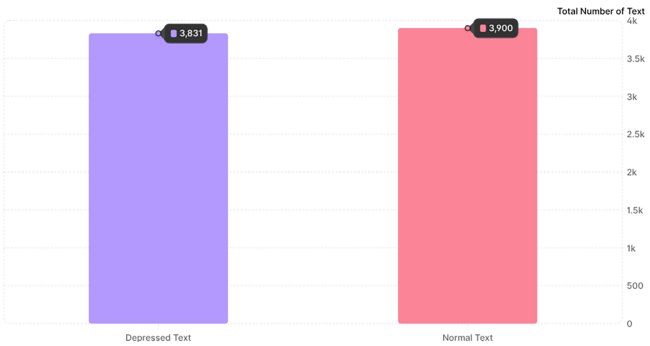

Figure 3: Dataset used for DepInferAttack model (Dataset 1).

To prepare the data for classification, we applied a preprocessing pipeline that included removal of duplicate posts, lowercasing, punctuation stripping, stopword removal, tokenization, and lemmatization. These steps helped reduce sparsity and standardize the textual input. The resulting dataset was used to train CNN-based text classifiers in the MIA experiments. Table 2 shows example posts from each class, highlighting the lexical and contextual variance exploited by MIA.

Table 2: Dataset 1 (DepInferAttack) sample posts.

Label	Sample Post
Normal	"Had a really productive day at work and finally finished my pending reports!"
	"Went for a long run this morning, feeling fresh and energized."
	"Just spent the weekend with family feeling grateful and happy."
Depressed	"I don't see the point in anything anymore. I'm tired all the time."
	"Been feeling empty lately. I can't talk to anyone about it."
	"Another day pretending to be fine when everything inside feels broken."

4.1.2 Dataset 2: BotInferAttack

The second experimental scenario is adapted from our previous work, BotInferAttack, where membership inference threats were demonstrated in the context of bot detection on online social networks. The dataset used in this scenario is derived from the publicly available Instagram dataset introduced by Akyon and Kalfaoglu [35], which contains 1400 labeled accounts, including 700 bot and 700 human profiles. This dataset was shown to be vulnerable to inference attacks due to distinguishable behavioral features encoded in the model outputs. Fig. 4 presents a bar chart showing the number of instances of both legitimate user and bot user used in the BotInferAttack Model.

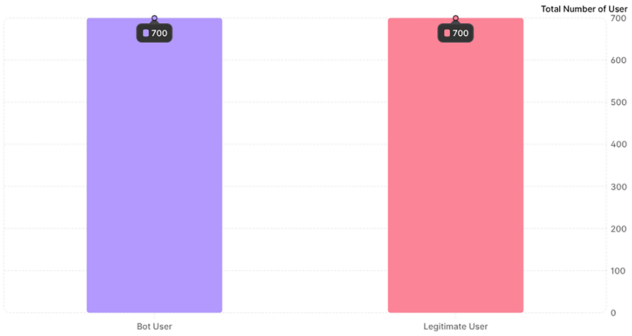

Figure 4: Dataset used for BotInferAttack model (Dataset 2).

The dataset includes both numerical and binary features that represent user activity patterns and profile characteristics. Specifically, features such as userMediaCount, mediaLikeNumbers_mean, userFollowerCount, and userFollowingCount capture quantitative behavioral signals. Binary indicators like userHasHighlightReels and userHasExternalUrl reflect user profile attributes, while features such as userTagsCount and userBiographyLength provide content-related metadata. Table 3 summarizes the complete list of features and their data types.

Table 3: Dataset 2 (BotInferAttack) feature details.

Feature Name	Feature Description
userMediaCount	The number of media a user has in their timeline.
mediaLikeNumbers_mean	The average number of likes received across all media posts shared by a user on their timeline.
userFollowerCount	The number of accounts that are following the user.
userFollowingCount	The number of accounts that are followed by the user.
userHasHighlighReels	Whether the user has highlighted any reel in their wall.
userHasExternalUrl	Whether the user has any external url in their wall.
userTagsCount	The number of tags user has been tagged.
userBiographyLength	The user's biography's length.
usernameLength	The length of the user's username.

All features were normalized using Min-Max scaling prior to model training. In our prior attack study, this dataset was used to successfully mount MIA against a neural classifier, revealing a critical need for effective post-inference defenses such as HEbdMIA. Both the dataset was divided into 80% training and 20% testing sets.

4.2 Model Training

Each dataset was used to train a separate target model that simulates the typical behavior of a deployed classifier vulnerable to inference attacks. For the DepInferAttack dataset [1], a convolutional neural network (CNN) was employed. The architecture included an embedding layer, followed by a convolutional layer, max-pooling, and dense layers leading to a softmax output. This architecture was chosen based on its effectiveness in prior text classification studies. For the BotInferAttack dataset [2], we used a fully connected neural network consisting of two hidden layers with ReLU activations and a softmax output layer. The model was trained using binary cross-entropy loss and the Adam optimizer.

Both models were trained on 80% of the data and validated on the remaining 20%. The bot detection model achieved accuracy of 94.2%, while the depression detection model achieved 90.1%. These models served as the target classifiers for all MIA evaluations in both baseline (unprotected) and HEbdMIA-protected configurations.

4.3 Membership Inference Attack Configuration

To evaluate model vulnerability, we adopt a shadow model-based attack framework following Shokri et al. [3], where an adversary trains a surrogate model to generate data for learning membership inference. Consistent with DepInferAttack [1] and BotInferAttack [2], we consider confidence, entropy, and logit-based leakage using three attack strategies: confidence thresholding, entropy-based inference, and a logistic regression classifier on logits. Each attack is performed on both original and HEbdMIA-protected outputs to measure changes in attack success rate. The evaluation focuses on black-box settings, reflecting practical deployment scenarios where internal model details are unavailable.

5 Result

This section presents the empirical evaluation of the proposed HEbdMIA framework in terms of its effectiveness against membership inference attacks (MIAs), preservation of model accuracy, and computational efficiency. The experiments are conducted on two representative high-risk scenarios DepInferAttack [1] and BotInferAttack [2] where previously work established the vulnerability of target models to MIA. Here, HEbdMIA is applied to the same attack surfaces to validate its capacity to mitigate inference leakage without compromising utility or scalability.

5.1 Membership Inference Attack

The primary evaluation metric is the attack success rate, defined as the proportion of correctly inferred training set memberships by the adversary. Table 4 summarizes the results across three types of attacks confidence thresholding, entropy-based inference, and logistic regression over logits both before and after applying HEbdMIA.

The effectiveness of HEbdMIA can be understood through common attack strategies: confidence thresholding exploits high prediction confidence, entropy-based inference targets low uncertainty, and logit-based classifiers learn patterns from raw outputs. By encrypting logits post-inference, HEbdMIA obscures these signals while preserving prediction accuracy, effectively mitigating diverse attack vectors

without retraining. Results in Table 4 confirm this, showing that baseline attack success rates exceeding 80% are reduced by approximately 20% across both DepInferAttack [1] and BotInferAttack [2]. This consistent reduction demonstrates that logit-level encryption disrupts adversarial exploitation of confidence and distributional cues. Moreover, the approach generalizes across both structured and unstructured data domains, with a reasonable privacy–utility trade-off suitable for real-world, privacy-sensitive applications.

Table 4: Attack success rate (%) without and with HEbdMIA on DepInferAttack and BotInferAttack.

Dataset	Attack Type	Attack Success Rate without HEbdMIA	Attack Success Rate with HEbdMIA	Reduction
Dataset 1 DepInferAttack [1]	Confidence Thresholding	76.2%	55.9%	20.3%
	Entropy-Based Inference	77.8%	58.2%	19.6%
	Logit-Based Classifier	82.4%	63.1%	19.3%
Dataset 2 BotInferAttack [2]	Confidence Thresholding	81.7%	65.8%	21.9%
	Entropy-Based Inference	79.3%	60.5%	18.8%
	Logit-Based Classifier	84.1%	62.3%	21.8%

5.2 Evaluation Metrics

While attack success rate indicates vulnerability, it does not fully capture adversarial performance. Therefore, we use standard metrics accuracy, precision, recall, F1-score, false positive rate, and membership advantage to comprehensively evaluate classifier-based membership inference.

Table 5 presents the inference performance metrics for the DepInferAttack [1]. The baseline model, which exposes raw logits, allows the adversary to infer membership with high precision and confidence. In contrast, the application of HEbdMIA significantly degrades adversarial performance, reducing all evaluation metrics to levels close to random guessing.

Table 5: Membership inference metrics for DepInferAttack model.

Metric	DepInferAttack without HEbdMIA	DepInferAttack with HEbdMIA	Reduction
Accuracy	94.4%	65.1%	29.3%
Precision	95.0%	63.5%	31.5%
Recall	93.7%	71.1%	22.6%
F1-Score	94.3%	67.1%	27.2%
False Positive Rate	4.9%	40.0%	35.1%
Membership Advantage	74.9%	23.3%	51.6%

Fig. 5 illustrate the confusion matrix of the DepInferAttack model which shows a higher number of correctly identified member samples whithout HEbdMIA, indicating the attack's effectiveness in exploiting confidence patterns. After applying HEbdMIA, the distribution becomes more balanced, demonstrating reduced attack success and improved privacy protection.

To validate the generalizability of these findings, we conduct a similar multi-metric analysis on the BotInferAttack [2] model. As shown in Table 6, the HEbdMIA defense yields consistent reductions in adversarial performance across all evaluation metrics. The precision, recall, and AUC drop sharply after encryption is applied, further demonstrating the defense's robustness across heterogeneous input modalities.

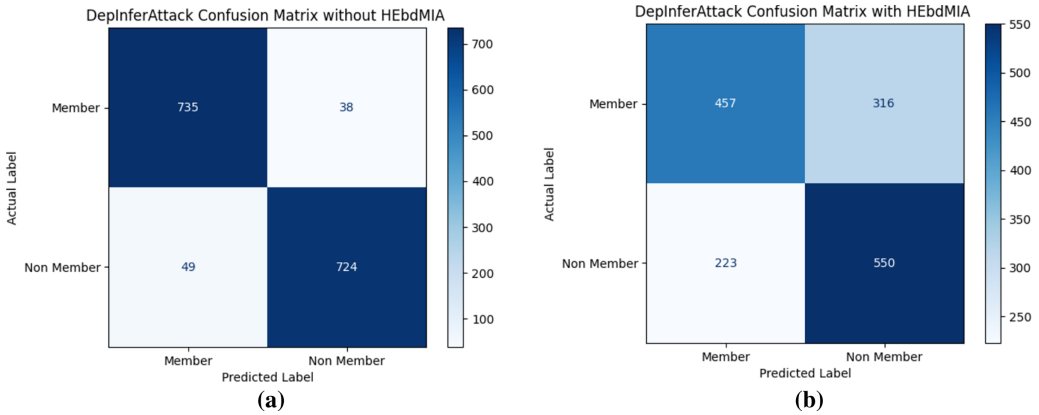

Figure 5: Confusion matrix analysis of DepInferAttack: (**a**) Without HEbdMIA (**b**) With HEbdMIA.

Table 6: Membership inference metrics for BotInferAttack model.

Metric	BotInferAttack without HEbdMIA	BotInferAttack with HEbdMIA	Reduction
Accuracy	90.0%	63.6%	26.4%
Precision	90.0%	64.0%	26.0%
Recall	90.0%	62.1%	27.9%
F1-Score	90.0%	63.0%	27.0%
False Positive Rate	10.0%	35.3%	25.3%
Membership Advantage	80.0%	27.1%	52.9%

Fig. 6 shows that, without HEbdMIA, the BotInferAttack model accurately identifies member samples, indicating strong attack effectiveness. After applying HEbdMIA, the confusion matrix becomes more balanced, reflecting reduced inference capability. Similarly, Tables 5 and 6 and Figs. 5 and 6 demonstrate a significant decline in attack performance for both DepInferAttack [1] and BotInferAttack [2], with decreases in accuracy, precision, recall, and F1-score, along with increased false positive rates. The reduction in membership advantage further confirms the effectiveness of HEbdMIA in weakening adversarial inference.

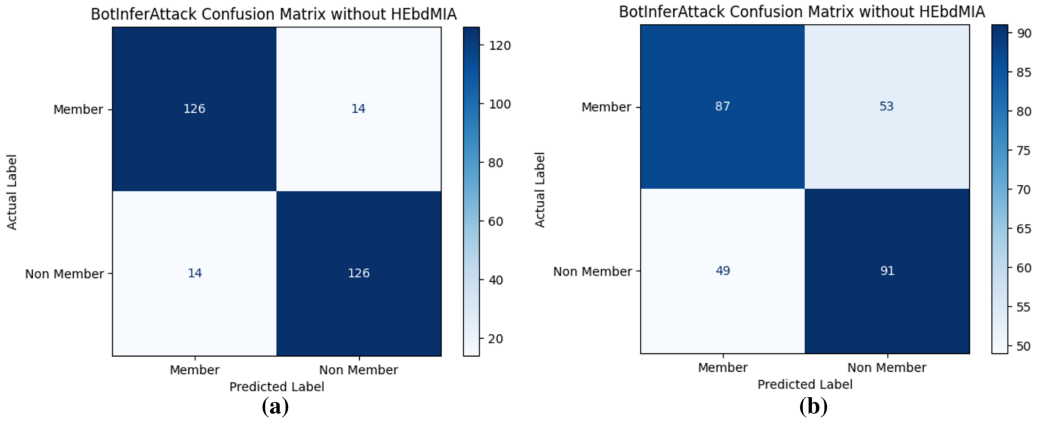

Figure 6: Confusion matrix analysis of BotInferAttack: (**a**) Without HEbdMIA (**b**) With HEbdMIA.

The ROC curve is a key metric for evaluating privacy defenses, as it measures an attacker's ability to distinguish member from non-member samples across all thresholds. Unlike accuracy, ROC-AUC provides a threshold-independent assessment of inference risk. As shown in Fig. 7, HEbdMIA reduces the AUC by approximately 30% for both DepInferAttack [1] and BotInferAttack [2], indicating a significant decline in adversarial discriminative power. This demonstrates that logit encryption effectively suppresses inference signals while maintaining practical model utility.

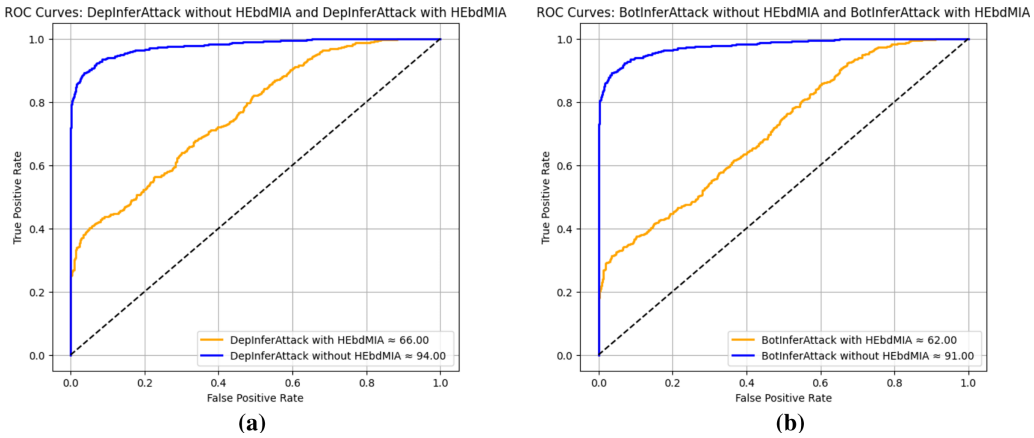

Figure 7: ROC curve for (**a**) DepInferAttack and (**b**) BotInferAttack model with and without HEbdMIA.

5.3 Computational Efficiency and Lightweight Design

Unlike fully homomorphic inference systems that apply encryption to model weights, inputs, or the entire computation graph, HEbdMIA operates exclusively on the model's output logits. Let the output of the trained model for an input sample x be a logit vector as:

$$\mathbf{z} = f(x) = [z_1, z_2, \ldots, z_k] \in \mathbb{R}^k \tag{6}$$

where k denotes the number of output classes. Each logit element is encrypted independently using the homomorphic encryption function defined in Eq. (2). Since the encryption function is applied individually to each element of the logit vector, the total encryption cost can be expressed as the summation of the encryption operations over all logits as:

$$T_{enc} = \sum_{i=1}^{k} C_{enc}(z_i) \tag{7}$$

where $C_{enc}(z_i)$ represents the computational cost of encrypting a single logit value. Because the Paillier encryption operation has a constant computational cost for each input value, we can denote

$$C_{enc}(z_i) = c \tag{8}$$

where c is a constant determined by modular exponentiation operations. Substituting this into the previous equation gives

$$T_{enc} = \sum_{i=1}^{k} c \tag{9}$$

which simplifies to

$$T_{enc} = c \cdot k \tag{10}$$

Similarly, the decryption of the encrypted logits performed by the trusted client follows

$$T_{dec} = \sum_{i=1}^{k} C_{dec}(z_i) \tag{11}$$

Assuming constant-time decryption for each ciphertext,

$$T_{dec} = c' \cdot k \tag{12}$$

where c' is the constant cost of the decryption operation. Therefore, the total computational overhead introduced by HEbdMIA can be expressed as

$$T_{total} = T_{enc} + T_{dec} = (c + c')k \tag{13}$$

Since c and c' are constants independent of the model architecture, dataset size, or input dimensionality, the overall time complexity reduces to

$$\mathbf{T_{HEbdMIA}} = O(k) \tag{14}$$

The linear $O(k)$ complexity arises from encrypting only the output logits, independent of model depth or input size. Since $k$is typically small (2–100), the computational overhead and latency remain minimal. This lightweight design avoids retraining or architectural changes, enabling seamless integration into existing machine learning pipelines.

The trade-off between privacy and utility is crucial in evaluating defense mechanisms. HEbdMIA achieves a reduction in MIA success rates of 31.0% (DepInferAttack [1]) and 27.3% (BotInferAttack [2]), with an accuracy decrease of 29.4% and 26.3%, respectively, demonstrating effective privacy protection with acceptable utility loss. From a computational perspective, conventional homomorphic encryption approaches incur significant overhead [36], whereas HEbdMIA operates only on post-inference logits, resulting in a linear complexity of $O(k)$ and minimal runtime and memory overhead. This lightweight design enables efficient deployment in real-time and resource-constrained environments. Overall, HEbdMIA provides a practical balance between privacy, accuracy, and efficiency, making it suitable for security-critical applications

5.4 Comparison With Existing Encryption-Based Defense Techniques

Encryption-based defenses such as HE-MAN [27], ReActHE [28], and SecureNN [30] mitigate membership inference attacks by securing model inputs and computations. However, HE-MAN [27] incurs high computational overhead, ReActHE [28] requires retraining and architectural changes, and SecureNN [30] introduces significant communication and computation costs due to secure multiparty protocols. Despite improved security, these approaches remain challenging to deploy in latency-sensitive and resource-constrained environments.

In contrast, the proposed HEbdMIA framework adopts a lightweight post-inference strategy by encrypting only the output logits, thereby eliminating the need for model retraining and enabling seamless integration into existing inference pipelines. Table 7 and Fig. 8 present a comparative analysis of encryption-based defense mechanisms against Membership Inference Attacks (MIA) on DepInferAttack [1] and BotInferAttack [2] models. The results indicate that existing approaches, such as HE-MAN [27],

ReActHE [28], and SecureNN [30] achieve only modest reductions in MIA success rates while imposing significant computational burdens. By comparison, HEbdMIA achieves a substantially higher reduction in MIA success rates 20.6% for DepInferAttack [1] and 21.3% for BotInferAttack [2] without requiring any model retraining. Furthermore, its time complexity of $O(k)$, dependent solely on the number of output classes, underscores its computational efficiency and suitability for real-time and edge-based deployments. Collectively, these findings demonstrate that HEbdMIA offers superior protection against membership inference attacks while preserving practical usability.

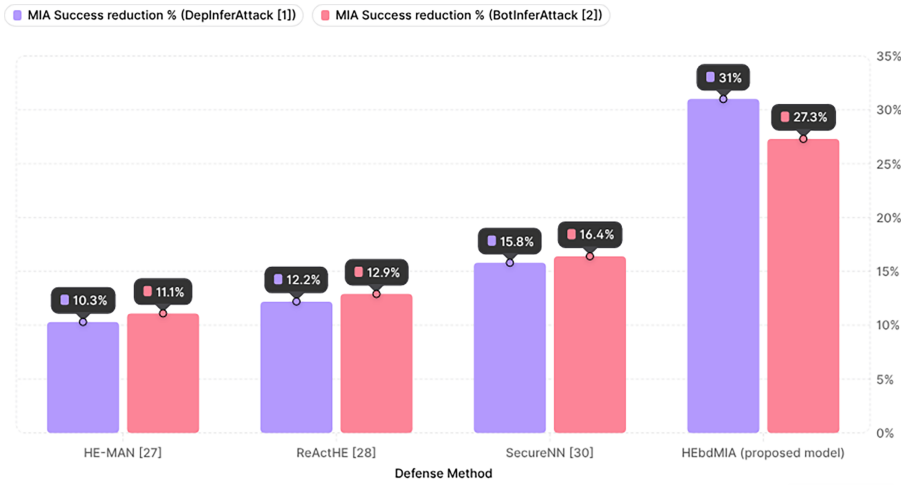

Figure 8: Comparison of MIA success drop rate for different encryption-based defense methods on DepInferAttack and BotInferAttack models.

Table 7: Comparison of encryption-based defense methods against membership inference attacks.

Defense Method	Encryption Scope	Retraining Required	MIA Success Reduction% (DepInferAttack [1])	MIA Success Reduction% (BotInferAttack [2])	Time Complexity
HE-MAN [27]	Full model (inputs + weights)	Yes	10.3%	11.1%	O(n2)
ReActHE [28]	Encrypted activation functions	Yes	12.2%	12.9%	O(2n)
SecureNN [30]	Model parameters + intermediate computations	Yes	15.8%	16.4%	O(n log n)
HEbdMIA (proposed model)	Output logits only (post-inference)	No	31.0%	27.3%	$O(k)$

5.5 Discussion

Membership Inference Attacks (MIAs) expose critical privacy vulnerabilities in modern machine learning systems, particularly in sensitive domains such as mental health analytics and social media profiling. Existing defenses, including adversarial regularization (e.g., MemGuard [18]), confidence masking, and homomorphic encryption-based approaches (e.g., HE-MAN [24], ReActHE [25]), either require retraining, degrade model utility, or introduce significant computational overhead. The proposed HEbdMIA framework addresses these limitations by applying lightweight post-inference logit encryption. Experimental results

demonstrate substantial reductions in MIA success rates 31.0% for DepInferAttack [1] and 27.3% for BotInferAttack [2] while incurring accuracy decreases of 29.3% and 26.4%, respectively. Although this trade-off is non-negligible, it is justified in privacy-sensitive applications where protecting user information is critical.

HEbdMIA preserves the relative ordering of predictions, ensuring that the model remains functionally useful while effectively disrupting the statistical patterns exploited by MIAs. Unlike prior methods that rely on adversarial training or computationally expensive encryption of the entire model, the proposed approach operates solely on output logits, achieving a computational complexity of $O(k)$, where $k$is the number of output classes. This design ensures scalability, as the method remains efficient regardless of model depth, input size, or dataset scale. Furthermore, the domain-agnostic nature of HEbdMIA validated across both textual (DepInferAttack [1]) and numerical/binary (BotInferAttack [2]) datasets demonstrates its generalizability. Overall, the proposed framework establishes that effective privacy protection can be achieved through lightweight, post-inference cryptographic transformation without compromising practical deployability.

6 Conclusion

Membership Inference Attacks (MIAs) pose a significant threat to the privacy and secure deployment of machine learning models, particularly in sensitive domains such as social media analytics and mental health assessment. In this work, we proposed HEbdMIA, a lightweight homomorphic encryption-based defense that operates at the post-inference stage by encrypting model output logits without requiring retraining or architectural modifications. Experimental results on DepInferAttack [1] and BotInferAttack [2] demonstrate that HEbdMIA achieves a substantial reduction in MIA success rates of 31.0% and 27.3%, respectively, while incurring an accuracy reduction of 29.3% and 26.4%, reflecting a controlled and interpretable privacy–utility trade-off. Additional evaluation using precision, recall, F1-score, and ROC-AUC further confirms a significant decline in adversarial inference capability. The proposed approach is model-agnostic, computationally efficient $O(k)$, and suitable for real-world deployment. Future research will focus on extending the framework to white-box and adaptive attack scenarios, enhancing scalability for large-scale models and complex architectures, and exploring integration with hybrid privacy-preserving techniques such as differential privacy and federated learning to further strengthen robustness and applicability.

Acknowledgement: During this work, the authors used QuillBot for language improvement and grammar checking, and take full responsibility for the final content.

Funding Statement: The authors would like to acknowledge the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R757), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would also like to thank Imam Mohammad Ibn Saud Islamic University (IMSIU) for their support.

Author Contributions: Conceptualization, methodology, and original draft editing: Akash Shah; supervision: Mudasir Ahmad Wani, Shri Kant, Ravi Prakash Chaturvedi, Nidhi Sindhwani; funding: Mudasir Ahmad Wani, Kashish Ara Shakil, Sulieman Alshuhri. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the First Author [Akash Shah], upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shah A, Varshney S, Mehrotra M. DepInferAttack: framework for membership inference attack in depression dataset. In: Proceedings of the 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS); 2024 Nov 13–15; Tashkent, Uzbekistan. doi:10.1109/ICTACS62700.2024.10840770.
2. Shah A, Mehrotra M, Kant S. BotInferAttack: an LSTM-based framework for membership inference attack on bot dataset. In: Proceedings of the 2025 2nd Global AI Summit-International Conference on Artificial Intelligence and Emerging Technology (AI Summit); 2025 Nov 19–21; Noida, India. doi:10.1109/AISummit66170.2025.11411003.
3. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP); 2017 May 22–26; San Jose, CA, USA. doi:10.1109/SP.2017.41.
4. Liu L, Wang L, Liu G, Peng K, Wang C. Membership inference attacks against machine learning models via prediction sensitivity. IEEE Trans Dependable Secur Comput. 2023;20(3):2341–7. doi:10.1109/TDSC.2022.3180828.
5. Shateri M, Messina F, Labeau F, Piantanida P. Preserving privacy in GANs against membership inference attack. IEEE Trans Inf Forensics Secur. 2023;19:1728–43. doi:10.1109/TIFS.2023.3342654.
6. Ekramifard A, Amintoosi H, Seno SAH. Protection of the CGAN against membership inference attack using Differential Privacy. Int J Inf Secur. 2022;25(2):67. doi:10.1007/s10207-026-01230-4.
7. Babar FF, Khan S, Parkinson S. Defending federated learning against adversarial attacks: a systematic literature review. Cyber Secur Appl. 2026;4(3):100127. doi:10.1016/j.csa.2026.100127.
8. Hu H, Salcic Z, Sun L, Dobbie G, Yu PS, Zhang X. Membership inference attacks on machine learning: a survey. ACM Comput Surv. 2022;54(11s):1–37. doi:10.1145/3523273.
9. Niu J, Liu P, Zhu X, Shen K, Wang Y, Chi H, et al. A survey on membership inference attacks and defenses in machine learning. J Inf Intell. 2024;2(5):404–54. doi:10.1016/j.jiixd.2024.02.001.
10. Bai L, Hu H, Ye Q, Li H, Wang L, Xu J. Membership inference attacks and defenses in federated learning: a survey. ACM Comput Surv. 2025;57(4):1–35. doi:10.1145/3704633.
11. Zhang G, Liu B, Zhu T, Ding M, Zhou W. Label-only membership inference attacks and defenses in semantic segmentation models. IEEE Trans Dependable Secur Comput. 2023;20(2):1435–49. doi:10.1109/TDSC.2022.3154029.
12. Fu X, Wang X, Li Q, Liu J, Dai J, Han J. Unlocking generative priors: a new membership inference framework for diffusion models. IEEE Trans Inf Forensics Secur. 2025;20(8):4638–50. doi:10.1109/TIFS.2025.3560776.
13. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016 Oct 24–28; Vienna, Austria. doi:10.1145/2976749.2978318.
14. Truex S, Liu L, Gursoy ME, Yu L, Wei W. Demystifying membership inference attacks in machine learning as a service. IEEE Trans Serv Comput. 2021;14(6):2073–89. doi:10.1109/TSC.2019.2897554.
15. Wu H, Cao Y. Membership inference attacks on large-scale models: a survey. arXiv:2503.19338. 2025.
16. Fang X, Kim JE. Center-based relaxed learning against membership inference attacks. arXiv:2404.17674. 2024.
17. Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security; 2018 Oct 15–19; Toronto, ON, Canada. ACM; 2018. p. 634–46. doi:10.1145/3243734.3243855.
18. Jia J, Salem A, Backes M, Zhang Y, Gong NZ. MemGuard: defending against black-box membership inference attacks via adversarial examples. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security; 2019 Nov 11–15; London, UK. doi:10.1145/3319535.3363201.
19. Ogburn M, Turner C, Dahal P. Homomorphic encryption. Procedia Comput Sci. 2013;20:502–9. doi:10.1016/j.procs.2013.09.310.
20. Chen Z, Pattabiraman K. Overconfidence is a dangerous thing: mitigating membership inference attacks by enforcing less confident prediction. arXiv:2307.01610. 2023.
21. Li J, Li N, Ribeiro B. MIST: defending against membership inference attacks through membership-invariant subspace training. arXiv:2311.00919. 2023.

22. Zhang Z, Ma J, Ma X, Yang R, Wang X, Zhang J. Defending against membership inference attacks: RM Learning is all you need. Inf Sci. 2024;670:120636. doi:10.1016/j.ins.2024.120636.
23. Chen Y, Chen J, Weng Y, Chang C, Yu D, Lin G. AdaMixup: a dynamic defense framework for membership inference attack mitigation. arXiv:2501.02182. 2025.
24. Yadav A, Mane S. Mitigating black-box membership inference attack using metric mapping. Int J Inf Secur. 2025;24(5):214. doi:10.1007/s10207-025-01123-y.
25. Yang Z, Yan X, Chen G, Tian X. Mitigating membership inference attacks via generative denoising mechanisms. Mathematics. 2025;13(19):3070. doi:10.3390/math13193070.
26. Gentry C. A fully homomorphic encryption scheme [dissertation]. Stanford, CA, USA: Stanford University; 2009.
27. Nocker M, Drexel D, Rader M, Montuoro A, Schöttle P. HE-MAN-Homomorphically Encrypted MAchine learning with oNnx models. In: Proceedings of the 2023 8th International Conference on Machine Learning Technologies; 2023 Mar 10–12; Stockholm, Sweden. doi:10.1145/3589883.3589889.
28. Song C, Shi X. ReActHE: a homomorphic encryption friendly deep neural network for privacy-preserving biomedical prediction. Smart Health. 2024;32(4):100469. doi:10.1016/j.smhl.2024.100469.
29. Byun J, Choi Y, Lee J, Park S. Privacy-preserving inference resistant to model extraction attacks. Expert Syst Appl. 2024;256(3):124830. doi:10.1016/j.eswa.2024.124830.
30. Wagh S, Gupta D, Chandran N. SecureNN: efficient and private neural network training. IACR Cryptol ePrint Arch. 2018;2018:442.
31. Mohammed SJ, Taha DB. Paillier cryptosystem enhancement for Homomorphic Encryption technique. Multimed Tools Appl. 2024;83(8):22567–79. doi:10.1007/s11042-023-16301-0.
32. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput. 2024;16(1):45–74. doi:10.1007/s12559-023-10179-8.
33. Yan H, Li S, Wang Y, Zhang Y, Sharif K, Hu H, et al. Membership inference attacks against deep learning models via logits distribution. IEEE Trans Dependable Secure Comput. 2023;20(5):3799–808. doi:10.1109/TDSC.2022.3222880.
34. Li H, Yu L, He W. The impact of GDPR on global technology development. J Glob Inf Technol Manag. 2019;22(1):1–6. doi:10.1080/1097198x.2019.1569186.
35. Akyon FC, Kalfaoglu ME. Instagram fake and automated account detection. In: Proceedings of the 2019 Innovations in Intelligent Systems and Applications Conference (ASYU); 2019 Oct 31–Nov 2; Izmir, Turkey. doi:10.1109/asyu48272.2019.8946437.
36. Fang Y, Yang X, Zhu H, Xu W, Zheng Y, Liu X, et al. Enhancing paillier to fully homomorphic encryption with semi-honest TEE. Peer-to-Peer Netw Appl. 2024;17(5):3476–88. doi:10.1007/s12083-024-01752-5.