# meMIA: Multi-level Ensemble Membership Inference Attack

Najeeb Ullah, Muhammad Naveed Aman, *Senior Member, IEEE,* and Biplab Sikdar, *Senior Member, IEEE*.

***Abstract*—Leakage of private information in machine learning models can lead to breaches of confidentiality, identity theft, and unauthorized access to personal data. Ensuring the safe and trustworthy deployment of AI systems necessitates addressing privacy concerns to prevent unintentional disclosure and discrimination. One significant threat, Membership Inference (MI) attacks, exploits vulnerabilities in target learning models to determine if a given sample was part of the training set. However, the effectiveness of existing MI attacks is often limited by the number of classes in the dataset or the need for diverse multi-level adversarial features to exploit overfitted models. To enhance MI attack performance, we propose meMIA, a novel framework based on stacked ensemble learning. meMIA integrates embeddings from a neural network (NN) and a Long Short-Term Memory (LSTM) model, training a subsequent NN, termed the meta-model, on the concatenated embeddings. This method leverages the complementary strengths of NN and LSTM models; the LSTM captures order differences in confidence scores, while the NN discerns probability distribution differences between member and non-member samples.**

**We extensively evaluate meMIA on seven benchmark datasets, demonstrating that it surpasses current state-of-the-art MI attacks, achieving accuracy up to 94.6% and near-perfect recall. meMIA's superior performance, especially on datasets with fewer classes, underscores the urgent need for robust defenses against privacy attacks in machine learning, contributing to the safer and more ethical use of AI technologies.**

## Impact Statement

This paper contributes to data privacy in machine learning (ML) models. Leakage of private information in machine learning models can have significant consequences for both individuals and the society as a whole. Although the data used for training ML models is kept private, unintentional disclosure of sensitive information through these models can result in breaches of confidentiality, identity theft, and unauthorized access to personal data. Furthermore, the likelihood of producing discriminating results rises when private data becomes available, intensifying issues regarding fairness and ethical considerations in the field of machine learning. Thus, the safe and trustworthy deployment of AI systems in diverse fields is contingent upon understanding privacy concerns in ML models. This paper presents a membership inference (MI) attack that takes advantage of a vulnerability in the target learning models to determine whether a given sample was a part of the training set or not. The existing MI attacks proposed in literature are limited either by the number of classes in a dataset or by the availability of various multi-level adversarial features to exploit overfitted target models. We show that by employing stacked ensemble learning with a neural network (NN) and LSTM, we can improve the accuracy of a MI attack to above 94% without making any restrictive assumptions on the number of classes or overfitting in the target model. These finding would help researchers in further improving the privacy of ML models.

***Index Terms*—Membership Inference Attacks, Privacy Leakage, Adversarial Attacks, Data Leakage, Recurrent Neural Network**

## I. Introduction

With the rise of digital technologies, privacy and security concerns have intensified. The use of machine learning models for decision-making is becoming ubiquitous in several sectors, including healthcare, finance, retail, education, social media, among others. However, it has been shown that ML models are vulnerable to adversarial attacks [18, 20, 32, 35, 36, 42, 41, 45, 46, 48, 61] that allow adversaries to infer information from a target machine learning model. Such attacks enable adversaries to find patterns in the output signals of the target model, leading to a comprehensive analysis of the private information used in training the target model.

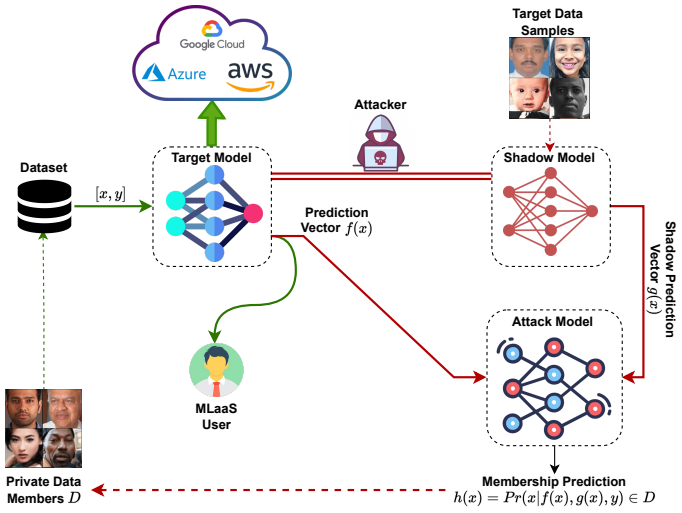


Fig. 1: Membership inference attack overview.

Machine learning models have been increasingly used to train

This work was supported in part by the Asian Institute of Digital Finance (AIDF) under grant A-0003504-09-00.

N. Ullah, and B. Sikdar are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, e-mail: {e1144142, bsikdar}@nus.edu.sg

M. N. Aman is with the School of Computing, University of Nebraska-Lincoln, Nebraska, USA, e-mail:naveed.aman@unl.edu

TABLE I: Summary of membership inference attacks (MIAs).

| Modality | Reference | Technique | Pros | Cons |
|---|---|---|---|---|
| **Text** | [2] S. Mahloujifar, et al. | Word Embeddings | Effective for text data, enhances model interpretability beyond just word embeddings | Resource intensive, limits the applicability of some findings to other types of models and data representations |
| | [47] A. Suri, et al. | Federated Learning | Has direct implications for privacy-preserving in healthcare and finance. | Lack of comprehensive defensive strategies could limit its practical utility for securing federated learning systems. |
| | [39] S. Rezaei, X. Liu | Discernibility Analysis | Offers a balanced perspective, challenging the perceived severity of MI attacks and encourages a more nuanced understanding of their implications. | May not provide complete solutions, theoretical focus |
| | [51] Y. Wu, et al. | Text-to-Image Models | Introduces innovative attack techniques, highlighting the vulnerabilities in generative models | Limits the generalizability to other types of generative models |
| **Images** | [3] R. Webster, et al. | GANs | Introduces effective attack methodologies and provides empirical evidence of vulnerabilities of GANs | Identifies vulnerabilities without proposing robust defensive mechanisms to secure GAN-generated content |
| | [62] J. Zhou, et al. | GANs | Expands understanding of privacy risks with generative models. | Limited to certain types of data, complex to implement |
| **Audio** | [56] Y. Yang, et al. | RNNs | Expanding the understanding of privacy risks in sequential models | Require significant resources, complex analysis |
| **Graph** | [19] X. He, et al. | Graph Neural Networks | Good for relational data, Introduces node-level MIA against GNNs | May not scale well to other types of NNs. |
| **General** | [23] H. Hu, J. Pang | Diffusion Models | Explores a relatively new area of diffusion models and underscores significant privacy and security implications | Computationally intensive, may require significant training data |
| | [4] A. Bagmar, et al. | Lottery Ticket Networks | Resource-efficient, leverages sparsity and provides valuable insights into the privacy risks associated with this model pruning technique. | May not scale well with other types of data. |
| | [22] D. Hu | Latent Factor Models | Effective for recommendation systems, scalable | Requires extensive data, model-specific vulnerabilities |
| | [14] M. Conti, et al. | Multi-Model Attacks | High attack success rate, covers multiple models | High computational cost, complex to defend against |
| | [54] M. Xu, X.-Y. Li | Data Origin Inference | This article proposed the novel concept of data origin inference attacks, providing valuable insights into data privacy issues beyond data membership. | Requires additional data, Computationally resource intensive |
| | [27] J. Liu, X. Lyu | Split Learning | Effective against distributed data in collaborative learning algorithms | Potential communication delays in practical settings, specificity to Split Learning |
| | [49] Y. Wang, et al. | Knowledge Graphs | Effective against structured data, formalizes membership inference attacks on knowledge graphs | May require large datasets |

on a wide range of sensitive data using MLaaS platforms.[1,2,3] Although the training platforms are trusted, for data owners, a significant worry remains: Can the model's output ( i.e., prediction vectors) be exploited to risk the confidentiality of sensitive training data? The complexity of machine learning models may obscure the risk of information leakage, but they can still reveal details about the training data. In particular, even in the most challenging black box setting, a membership inference attack (MIA) [21, 43] on a model can be carried out by an adversary to determine if a specific data record was included in the training set [45] or not. In MI attacks, the adversary takes advantage of the distinct response of the target model on training and testing datasets, posing data privacy concerns.

Existing attack methods [45, 57, 31, 34] in the black-box setting often rely on a limited amount of information per sample for training the attack model. In particular, shadow model-based approaches commonly rely on prediction vectors to discern between member and non-member samples. Despite

[1]https://cloud.google.com/vertex-ai
[2]https://aws.amazon.com/machine-learning
[3]https://azure.microsoft.com/en-us/products/machine-learning

their utility, these approaches exhibit limitations in specific scenarios. When faced with a limited number of classes per dataset, the prediction vector does not provide enough information to distinguish between member and non-member samples accurately. Moreover, when the number of samples per class is limited, the attack model struggles to learn the complex boundaries of each class, leading to a significant drop in performance.

Recently, an alternative method, seqMIA [28] treats prediction vectors as sequences and trains Recurrent Neural Networks (RNNs) to learn the order within confidence scores. This approach assumes that member and non-member samples exhibit different orderings in their confidence scores. Despite seqMIA's superior inference performance, it does not fully utilize hidden information in the prediction vectors.

We argue that limiting an attack model's training to a single level of knowledge, like probability distribution or order differences, is insufficient for optimal inference accuracy. Therefore, we introduce a novel attack method leveraging multi-level information gleaned from two distinct networks. The first neural network (NN) distinguishes member from non-member samples by learning their contrasting probability distributions. Simultaneously, a recurrent neural network with Long Short-Term Memory (RNN-LSTM) captures the underlying rank (order) differences in confidence scores between members and non-members. This rich multi-level information is then harnessed by training the final attack model on the combined embedding representations generated by the two types of neural networks, empowering it to achieve improved attack performance. This multi-level ensemble membership inference attack (meMIA) approach enables the attack model to capture richer features and significantly enhance its membership inference capabilities. We use RNN-LSTM to capture the priori order differences in our attack model because it performs well compared to vanilla RNN and GRU, as shown by [28].

In Table I, we summarize different techniques, their advantages, and limitations across multiple modalities, such as text, images, audio, and graphs, to provide a comprehensive comparative analysis of current MIAs. We observe that most current attack methods require a significant amount of processing power and are sensitive to the hyperparameters of the target models. They also rely on access to large datasets and struggle to generalize well. Additionally, these methods display specificity to certain models and have limited use of adversarial knowledge. In contrast, meMIA utilizes a unique framework to address these issues by targeting models trained on images in a black-box environment using the confidence scores of the shadow model.

Our proposed method, as shown in Figure 2, demonstrates several notable advantages: Firstly, leveraging multi-level information enables the attack model to achieve significantly higher accuracy than state-of-the-art MIA methods, especially in scenarios with limited class sizes or restricted samples per class. This improved accuracy translates to a more robust and reliable attack. Furthermore, the proposed approach exhibits enhanced flexibility, readily adapting to different prediction vectors and seamlessly integrating into diverse MI attack scenarios. Finally, the multi-level embedding generates a unified representation that encapsulates distribution and order differences, contributing to the overall improvement in attack performance. This unified representation provides a powerful tool for understanding and exploiting the underlying discrepancies between member and non-member samples, ultimately leading to more effective MIA attacks in challenging black-box settings.

The major contribution of this paper are as follows:

- We introduce a novel attack method that utilizes multi-level adversarial knowledge obtained from primary ensemble learning models, offering improved performance compared to state-of-the-art attack methods.
- A comprehensive analysis of MI attack performance is presented, which compares meMIA with state-of-the-art MI attack methods, using multi-model datasets with varying levels of complexity.
- We present a stacked ensemble technique that leverages established adversarial features like probability order difference and probability distribution difference to differentiate between data members and non-members. This effectively makes meMIA scalable and adaptable to any *newly found adversarial knowledge* by adding a base NN model (dedicated to newly found adversarial knowledge) that helps jointly increase the learning capacity of the meta-model even more
- meMIA excels in providing superior attack inference performance for datasets with fewer classes.

The rest of the paper is structured as follows. Section II discusses machine learning background and privacy implications concluding on black-box attack setting. In Section III, we introduce benchmark datasets, target model used to evaluate meMIA's performance. The details of the proposed meMIA scheme are discussed in Section IV, followed by a brief description of the experimental setup in Section V. A comprehensive meMIA performance analysis is provided in section VI. Finally, Section VII concludes the paper.

## II. PRIVACY IN MACHINE LEARNING

Machine learning algorithms help AI learn without direct programming for specific actions. These algorithms identify patterns in complex datasets, enabling AI to make predictions and execute tasks based on these patterns rather than pre-written instructions. In supervised machine learning, each sample is input into the model along with its corresponding target label or ground truth. The essence of this process lies in deciphering the connections between the dataset and its associated target labels, with an objective to develop a model capable of effectively applying these insights to data samples not included in the training set [17]. On the other hand, in unsupervised machine learning, the primary aim is to discover valuable attributes from data without predefined target labels, thereby constructing a model that reveals the underlying patterns from unlabeled datasets. Furthermore, machine learning models tend to overfit data while minimizing prediction error. To address this problem, various regularization techniques have been introduced, aiming to balance the reduction of prediction errors while ensuring the model does not excessively conform to the specificities of its training data [17, 34]

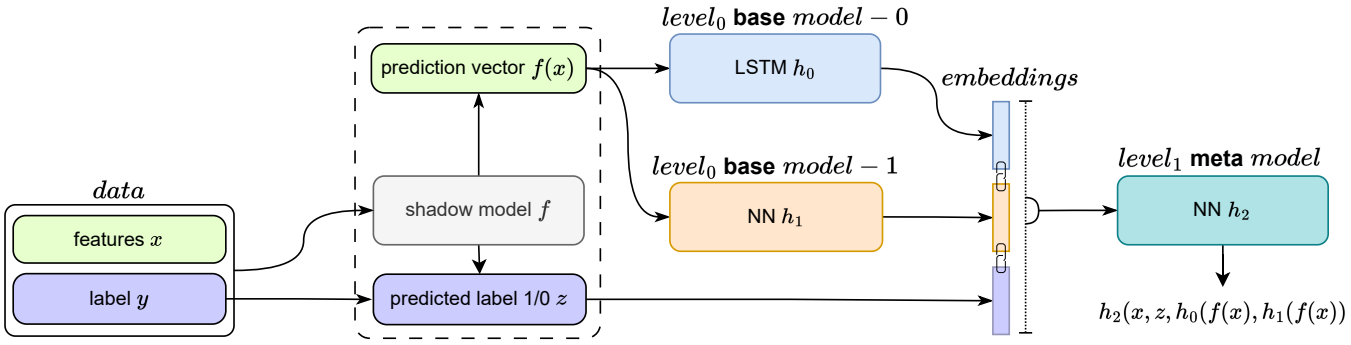


Fig. 2: Proposed inference attack model, the meMIA

Supervised machine learning techniques have been used in several applications; for instance, a retail company might develop a system to anticipate a customer's purchasing style, offering tailored incentives. Similarly, a healthcare provider could employ a model for determining the most effective treatment based on a patient's symptoms or genetic profile. Due to the ubiquity of ML as a tool, the end user may be a novice in designing, developing, training, and deploying large-scale models. Therefore, several tech firms now offer machine learning as a service (MLaaS) on their cloud platforms. A typical use case of MLaaS might involve an app developer collecting user data, using MLaaS to train or update a model, and then applying the model's insights to enhance the app's functionality or user interaction. However, the details of these models and their training processes are not disclosed to the users. The service may automatically select the model type based on the data and validation accuracy. These platforms often fail to inform users about risks like overfitting and provide limited control over regularization (e.g., Google Prediction API conceals most details, while Amazon ML offers only basic options like L1 or L2 regularization). Recent studies [45, 7] have shown significant data privacy and security concerns with these MLaaS platforms, demonstrating considerable data leakage risks.

Next, we present four inference attack methods and explore related work in privacy-preserving machine learning that aims to protect against these attacks, with a particular emphasis on membership inference attacks.

### A. Model Stealing

The objective for an adversary in these attacks is to replicate the target model effectively, creating a "stolen" model [35, 48]. In essence, the attacker intends to get a model with similar inference performance compared to the original model. For instance, [35] demonstrated how an attacker could use queries to a target model to reconstruct a high-fidelity copy, highlighting significant privacy risks in MLaaS. Similarly, [48] showed that even with black-box access, attackers could use a large number of queries to train a surrogate model to mimic the target model's behavior. In contrast, meMIA does not depend on a large number of queries. Instead, meMIA trains shadow model on a dataset that has the same format, size, and distribution. However, the training data samples for the shadow model do not overlap with the training samples of the target model as discussed in Section V.

### B. Attribute Inference

While training, the ML model not only learns to predict the primary label but also correlates other labels for the same sample. For instance, as demonstrated by Melis et al. [32], an ML model not only performs its main task, like predicting age, but also learns to predict attributes like race. These attack methods exploit such unintended information leakage. The success of an attribute inference attack is dependent on the assumption of having white-box access to the target model, meaning the attacker has complete knowledge of the model's structure and parameters. This reliance limits their efficacy in practical situations where the attacker might not always have white-box access. On the other hand, meMIA is primarily designed for black-box access attack settings, where the attacker has minimal to no knowledge about the target model. This makes meMIA more flexible and practical in real-world scenarios since it can operate without requiring detailed information about the target model.

### C. Model Inversion

Model inversion attacks aim to recreate the original training data, including the recovery of individual features for each sample, see [18] for detailed analysis. These attacks are only effective if the attacker has white-box access to the model. This is because the attacker needs to perform backpropagation passes on the model's parameters to reconstruct a data point. Moreover, the output of the model inversion is an average of the features that, at best, can categorize an output class. More precisely, it does not construct an identical data sample, nor can it identify if a specific data sample was used in the training. By contrast, meMIA's framework allows the attacker to operate in a black-box setting with optimal data membership identification capabilities. It uses stacked ensemble learning, combining NN and LSTM model embeddings to improve attack performance.

### D. Membership Inference

The membership inference attack method involves the adversary finding out whether a particular data sample has been

TABLE II: A description of different datasets used in the evaluation.

| Datasets | # of classes | Resolution | # Epochs (target model) | Training set (target model) | Training set (shadow model) | Target set |
|---|---|---|---|---|---|---|
| UTKFace | 4 | 32 × 32 | 80 | 5,500 | 5,500 | 11,000 |
| FMNIST | 10 | 32 × 32 | 80 | 16,280 | 16,280 | 32,560 |
| STL-10 | 10 | 32 × 32 | 80 | 3,250 | 3,250 | 6,500 |
| CIFAR-10 | 10 | 32 × 32 | 80 | 10,000 | 10,000 | 20,000 |
| Location | 30 | N/A | 60 | 1,252 | 12,52 | 2,504 |
| CIFAR-100 | 100 | 32 × 32 | 80 | 14,770 | 14,770 | 29,540 |
| Purchase-100 | 100 | N/A | 60 | 10,000 | 10,000 | 20,000 |

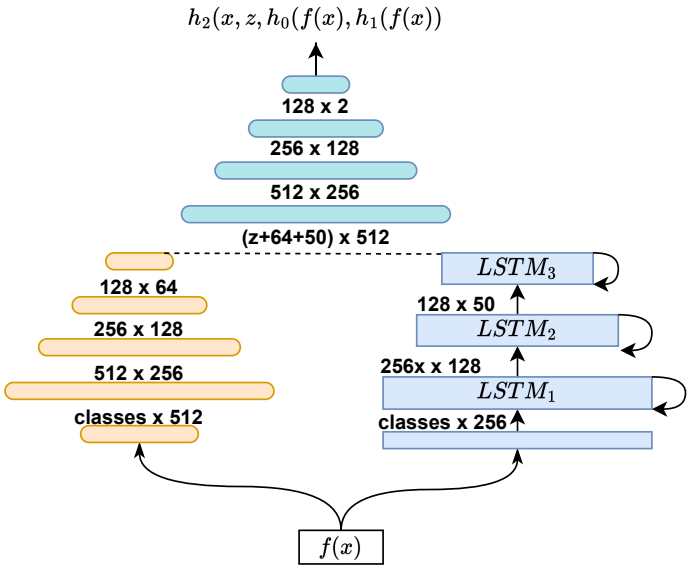


Fig. 3: meMIA with a more detailed architectural view.

used in the training. In other words, the attacker is interested in determining a data sample's membership (member, non-member). Usually, these attacks serve as gateways and, when combined with other attack methods, provide the first indication of whether the target model is vulnerable [15]. MI has been rigorously investigated in the literature [11, 12, 24, 25, 34, 41, 45]. Inferring membership of a target sample prompts severe privacy threats; for instance, if an ML model for drug dose prediction is trained using data from patients with a particular disease, then inclusion in the training set inherently leaks the individuals' health status. Moreover, we will discuss configurations that an adversary can use for MI attacks based on their accessibility to the target model.

*1) Black-Box/Shadow:* We begin with the most common and difficult attack scenario, as described by Salem et al. [41] and Shokri et al. [45]. In this setting, the adversary possesses an auxiliary dataset (the shadow dataset) and black-box access to the target model. The adversary divides the shadow dataset into two halves. One half trains the shadow model, and the other serves as a test set. Following the training of the shadow model, the adversary uses the entire shadow dataset to make queries to this model For each query, the shadow model returns posterior probabilities (prediction vectors) and corresponding labels. To prepare the data for the attack, the adversary assigns 'member' labels to the samples that were part of the shadow model's training set and 'non-member' labels to those used in the testing phase. The adversary then uses the labelled dataset to train an attack model, a binary classifier for membership inference. Once trained, the attack model uses the target model's posterior probabilities to predict whether the data sample was a member or used in training.

*2) Black-Box/Partial:* In this setting, the adversary has partial access to the actual data used in the target model's training, but they have black-box access to the target model itself [41]. Therefore, the attack method is similar to the Black-Box/Shadow approach. However, creating a shadow model is not mandatory for the attacker in this scenario. Instead, they can directly train their attack model using the partial training dataset as a reference point for identifying membership.

*3) White-Box/Shadow:* Nasr et al. [33] propose a white-box attack using a shadow or partial training dataset as an auxiliary resource. In scenarios similar to the black-box/shadow setting, the adversary employs a shadow auxiliary dataset to train a parallel model known as the shadow model. This shadow model is designed to imitate the target model's behaviour, thereby generating data to train the adversary's attack model. When it comes to a white-box setting, it is essential to note that the attacker has direct access to the target model. This access enables them to take advantage of different aspects of the target sample, such as its gradients in relation to the model parameters, embeddings from various intermediate layers, classification loss, and the prediction posteriors, including the label.

*4) White-Box/Partial:* The attack methods that use this setting are based on the same parameters, features, and gradients as target models, according to [33]. This method is the inverse of black-box/shadow. In conclusion, we focus on the Black-Box/Shadow setting, the most challenging setting to evaluate an MI attack.

### *E. Privacy-preserving machine learning*

Existing literature on privacy preservation in machine learning mainly focuses on enabling models to learn from data without compromising privacy. Henceforth, Secure Multiparty Computation (SMC) allows multiple parties to jointly compute a function over their inputs while keeping those inputs private using decision trees [26]. Moreover, in techniques such as k-means clustering [37] and Naive Bayes classifiers [60], the goal is to limit information leakage during training. Besides, The training algorithm used in privacy-preserving models remains unchanged from that of non-privacy-preserving models, meaning that these models are as susceptible to inference attacks as those trained by conventional methods. Therefore, the vulnerability is extended to models trained using computations on encrypted data [8, 53]. On the other hand, Differential

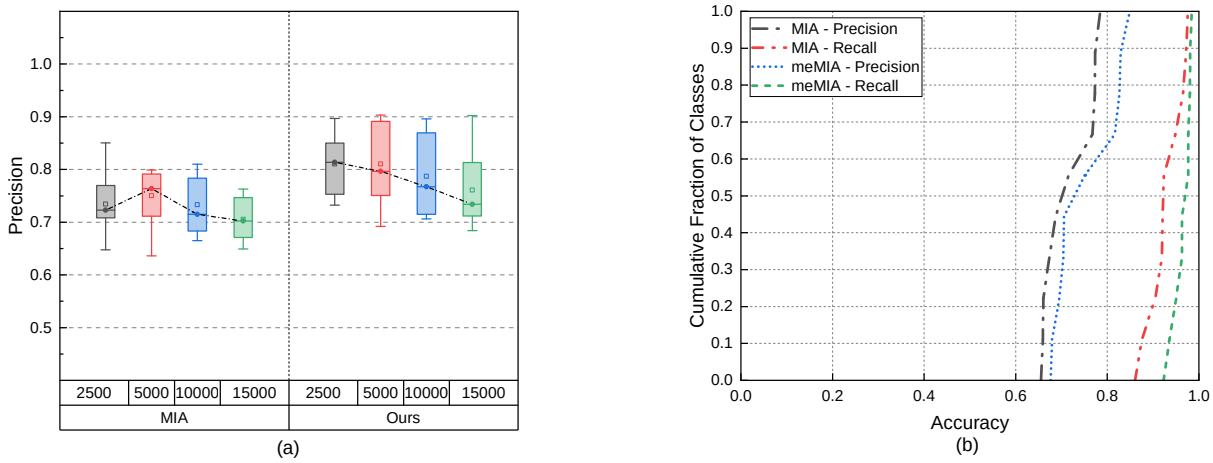


Fig. 4: Precision of the membership inference attack against CNN-based neural network trained on CIFAR-10 dataset. (a) shows precision for varying dataset sizes while comparing MIA and our attack methods. (b) presents the Empirical CDF (cumulative fraction of classes) of the precision and recall for MIA and Our membership inference attack, meMIA.

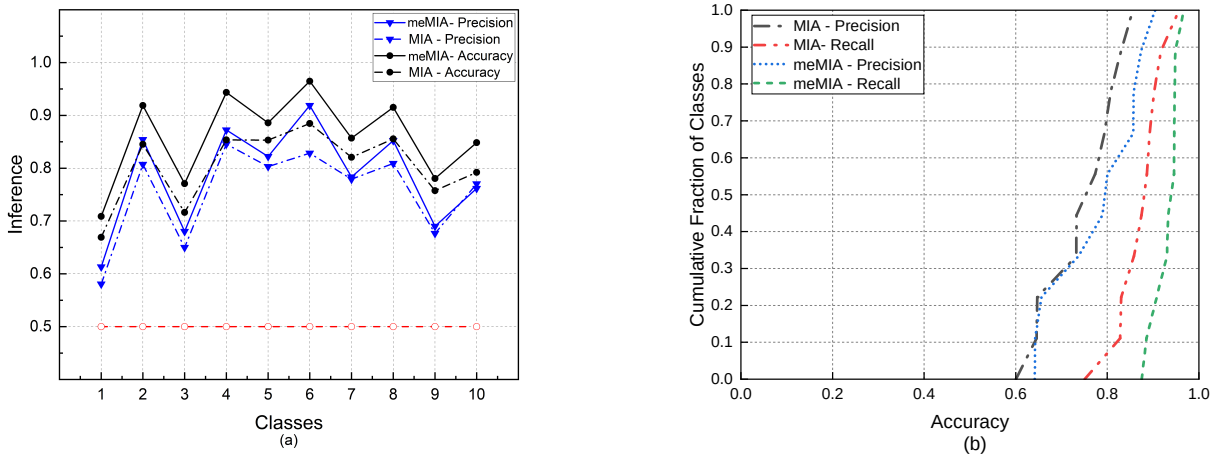


Fig. 5: Precision and recall of the membership inference against CNN-based neural networks on STL-10 dataset of size 6500. (a) compares MIA and our attack methods per class with random guessing 0.5. Moreover, (b) shows the Empirical CDF (cumulative fraction of classes) of the precision and recall for MIA and Ours membership inference attacks.

Privacy (DP) [16] has been widely accepted and implemented in various machine learning algorithms across different applications, which include support vector machines [40], linear and logistic regression [10, 59], and deep learning [6, 44]. DP reduces the likelihood of successful membership inference attacks by nature, which are predicated on extracting data specifics from the model outputs. Consequently, we believe that DP-based techniques also pose the same problem to our attack method as all SOTA MI attack methods [45]. Nonetheless, as elicited by [9], no matter how sophisticated the protection scheme is, the inference performance of an MI attack will always be above 50%, which is still alarming.

## III. PRELIMINARIES

In this section, we begin by defining the datasets used in our experiments, providing a foundation for the subsequent analyses. Following this, we briefly discuss the target models and then shift focus to a comprehensive description of our proposed attack model. In subsequent subsections, we discuss the experimental setup, followed by the results of the experiments.

### A. Datasets

In this paper, we use several benchmark datasets to evaluate meMIA. Additionally, to make the comparison fair among all image-based datasets, we resize them to 32x32 pixels.

*1) **FMNIST (Fashion-MNIST)**:* [52] comprises of 70,000 grayscale images, each with dimensions of 28x28 pixels. The images showcase various fashion items including t-shirts/tops, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. Notably, these fashion items are evenly distributed across 10 distinct classes within the dataset. We

randomly select 10,000 mutually exclusive samples for each shadow and target model training. The target and shadow datasets have 20,000 samples each, i.e., 10,000 for members and 10,000 for non-members.

*2) **UTKFace**:* [1] encompasses 23,000 face images annotated with age, gender, and race information. In this paper, we focus on images from the four racial categories (White, Black, Asian, and Indian) within the dataset, employing race as the label for the target models. As a result, we work with a subset of 22,012 images with 4 classes. For training, we randomly select 5,500 samples that are mutually exclusive for each shadow and target model. The target and shadow datasets both have 11,000 samples each, i.e., 5,500 for members and 5,500 for non-members.

*3) **Location**:* We use the processed version of the publicly accessible dataset containing location "check-ins" from mobile users on the Foursquare social network [55].[4] The processed samples consist of users with fewer than 25 check-ins and venues with fewer than 100 visits, resulting in a dataset comprising 5010 records used in prior work [45]. Each record is characterized by 446 binary features indicating whether the user has visited specific regions/locations. The dataset is clustered into 30 classes, each representing a geosocial type. For training, we randomly select 1,252 samples that are mutually exclusive for each shadow and target model. The target and shadow datasets both have 25,00 samples each, i.e., 1,252 for members and 1,252 for non-members.

*4) **Purchase-100**:* This dataset contains shopping histories from Kaggle's "Acquire Valued Shoppers" challenge.[5] In our experiments, we utilized the simplified version with 197,324 records clustered into 100 classes from [45], each representing a specific purchase style. Further, each class is represented by 600 binary features corresponding to a product and represents whether the user has purchased it or not. Given 600 binary features, the classification task is to predict the user's purchase style. For training, we randomly select 10,000 disjoint samples for each shadow and target model. The target and shadow datasets both have 20,000 samples each, i.e., 10,000 for members and 10,000 for non-members.

*5) **CIFAR**:* CIFAR-10 and CIFAR-100 [5] are famous benchmark datasets used to evaluate pattern recognition Machine learning models; both comprise 60,000 32x32 colour images equally distributed across 10 and 100 distinct classes, respectively. Both datasets are partitioned into 50,000 images for training and 10,000 for testing. To comprehensively investigate the impact of target training data size and compare the attack inference performance of meMIA to a baseline MI attack, we use different fractions of the CIFAR-10 dataset similar to [45] e.g., [2,500, 5,000, 10,000 and 15,000]. For CIFAR-100, we randomly select 14,770 disjoint samples for each shadow and target model. The target and shadow datasets have 29,540 samples each, i.e., 14,770 for members and 14,770 for non-members.

*6) **STL-10**:* [13] comprises 1,300 images within each class. The classes include objects and animals: aeroplanes, birds, cars, cats, deer, dogs, horses, monkeys, ships, and trucks. STL-10 classes show more diversity with fewer samples per class than CIFAR-10, introducing different difficulty levels for prediction models. We randomly select 3,250 disjoint samples for each shadow and target model training. The target and shadow datasets both have 6,500 samples each ,i.e., 3,250 for members and 3,250 for non-members.

[4]https://sites.google.com/site/yangdingqi/home/foursquare-dataset

[5]https://kaggle.com/c/acquire-valued-shoppers-challenge/data

TABLE III: The training and testing accuracies of the target models on seven benchmark datasets.

| Dataset | Target Model | Training Acc. | Testing Acc. |
|---|---|---|---|
| FMNIST | CNN | 0.999 | **0.900** |
| UTKFace | CNN | 0.999 | **0.832** |
| STL-10 | CNN | 0.999 | 0.548 |
| CIFAR-10 | CNN | 0.998 | 0.602 |
| CIFAR-100 | CNN | 0.997 | **0.272** |
| Location | NN | 0.989 | 0.687 |
| Purchase-100 | NN | 0.999 | 0.660 |

### B. Target Models

In this paper, we have implemented three distinct neural network architectures for different datasets. For instance, we use a simplistic CNN architecture for FMNIST, UTKFace, CIFAR-10, CIFAR-100, and STL-10 datasets. The CNN model was developed with three convolutional layers, each with a kernel size of 3. After the convolutional layers, the model uses max-pooling with a kernel size of 2 and ends with two fully connected layers as proposed by [30]. Conversely, a 2-layer fully connected neural network is employed for the Location dataset with layer sizes [128, 30]. For the Purchase-100 dataset, a 6-layer fully connected neural network is utilized with layer sizes of [2048, 1024, 512, 256, 100]. We use the ReLU activation function across all architectures. We follow prior work [45] to design our target model for image-based datasets and set a few hyperparameters in a similar way as described in [30]. Conversely, we calibrated the hyperparameters for non-image datasets to achieve a specific target testing accuracy, giving us precise control over the target model's generalizability. This approach allowed for a thorough assessment of the meMIA attack model's capabilities.

We set 64 as the mini-batch size for training the target model and chose cross entropy as our loss function. Stochastic Gradient Descent (SGD) was used as an optimizer, incorporating a weight decay of $5e-4$ and a momentum of 0.9 similar to [30]. We train target models for 60 epochs, with a learning rate of $1e-2$ for the initial 50 epochs and a subsequent adjustment to $1e-3$ from [50-100] epochs. Table III shows the training and testing accuracy of the respective target models. Additionally, we train shadow models in a similar fashion.

## IV. PROPOSED ATTACK MODEL

Stacked-ensemble learning has proved to be one of the most popular meta-learning techniques [50]. This method employs a meta-model to discern the reliability of various base models [58]. In this technique, the output of several base models is used as input to a meta-model, which makes the final prediction. The base models are trained using their respective data

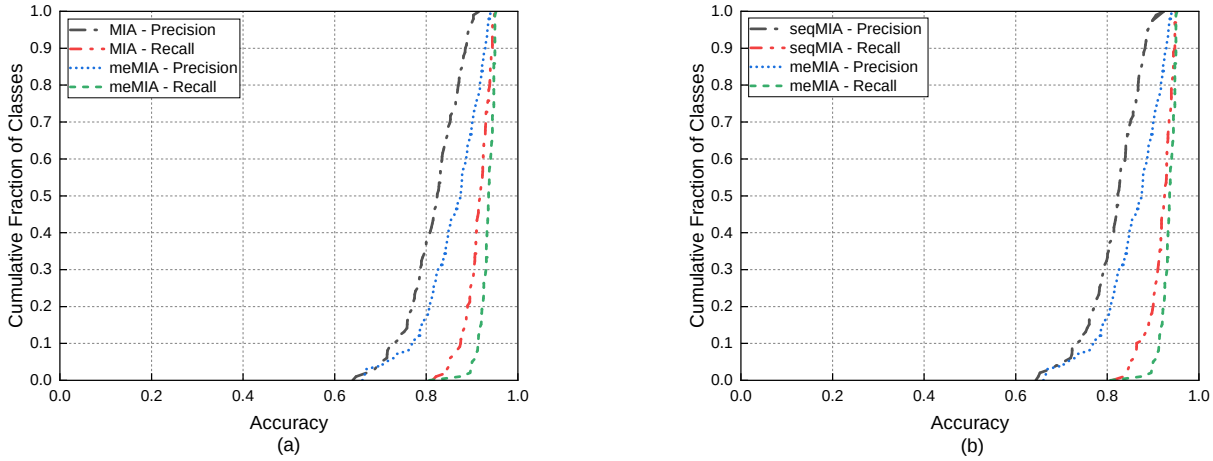


Fig. 6: Empirical CDF (cumulative fraction of classes) of the precision and recall for MIA and Ours membership inference attacks trained locally using neural networks for the CIFAR-100 dataset of size 29,540. (a) compares baseline MIA and our attack, and (b) shows a comparison between seqMIA and our attack method.

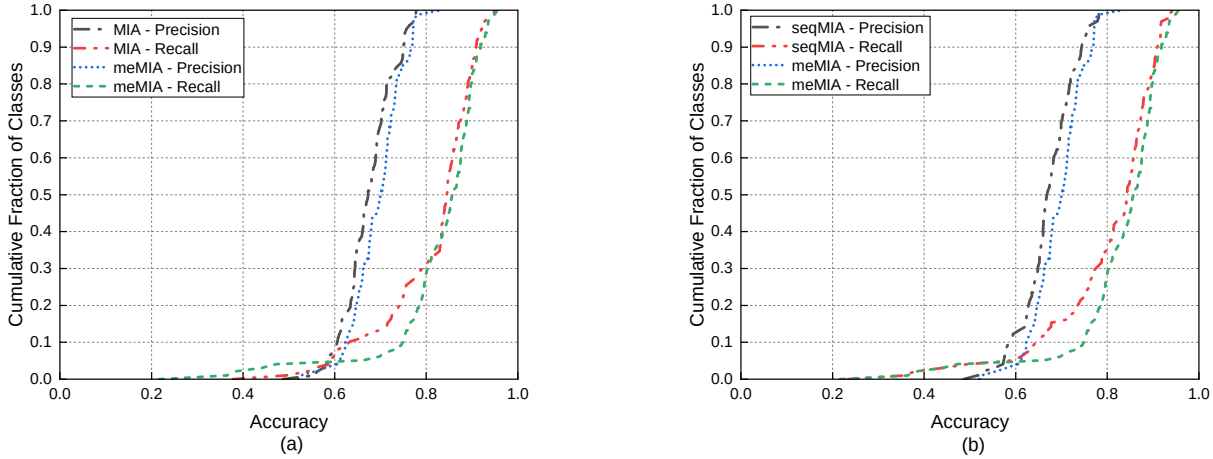


Fig. 7: Empirical CDF (cumulative fraction of classes) of the precision and recall for MIA and Ours membership inference attacks trained locally using neural networks for Purchase-100 dataset of size 20,000. (a) shows a comparison between the baseline MIA and our attack, while (b) shows a comparison between seqMIA and our attack method.

and produce the corresponding predictions. These predictions are then combined and used as inputs to the meta-model, which learns from them to make a final prediction as illustrated in Figure 9. Stacking allows combining the strengths of several models and producing a more accurate and robust prediction than any single model alone [38].

In the black box setting, training a model with a dataset that has fewer classes severely degrades the attack model inference performance. Moreover, researchers have shown increased that the inference attack performance increases as the number of classes increases [45]. Similarly, Nasr et al. [34] combines shadow prediction vector, one-hot encoded label, and shadow model prediction to train their attack model.

Our proposed attack model, meMIA, employs an ensemble of base models, comprising a simple NN, a multi-layer LSTM, and a meta-model that learns from the combined embeddings of base model as shown in Figure 2. meMIA is based on the key insight that two different neural networks learn independently from the same data sample, leading to diverse features. Until recently, we have not found any prior work that has observed this gap. All the existing attack methods train the attack model to capture the probability distribution difference or probability order difference to predict members and non-members. However, our attack method lets $level_1$ NN $meta{-}model$ combine the final output embeddings of the base models where $level_0$ NN $base - model_1$ captures the priori distribution difference and $level_0$ LSTM $base{-}model_0$ learns order differences in prediction vectors as illustrated in Figure 2, showing the architecture of our inference neural network and Figure 3 illustrating a more detailed view.

The base NN provides insights into the probability distribution differences, while the LSTM contributes understanding of

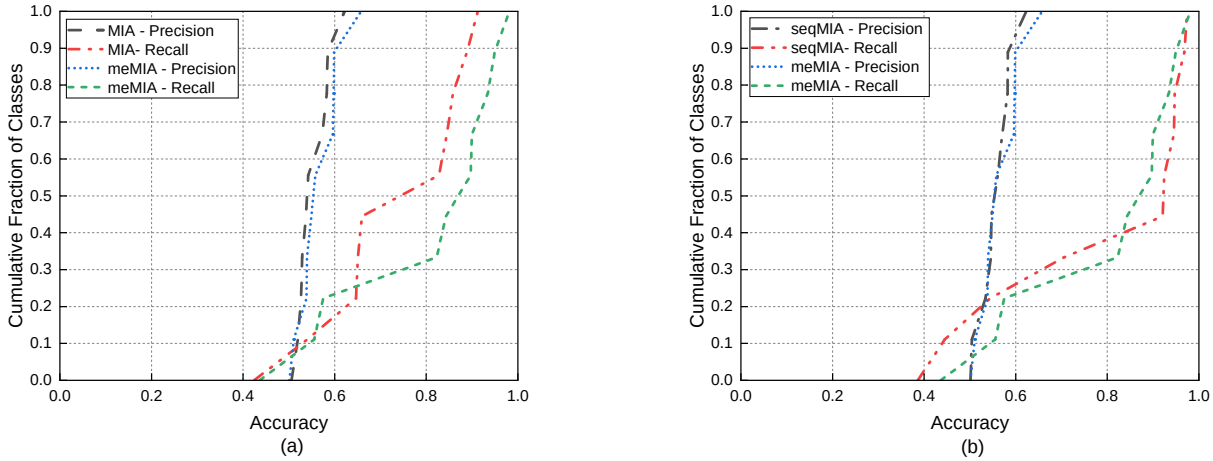


Fig. 8: Empirical CDF (cumulative fraction of classes) of the precision and recall for MIA and our attacks trained locally using CNN-based neural network for FMNIST dataset of size 20,000. (a) shows a comparison between baseline MIA and our attack, and (b) presents a comparison between the seqMIA and our attack method.

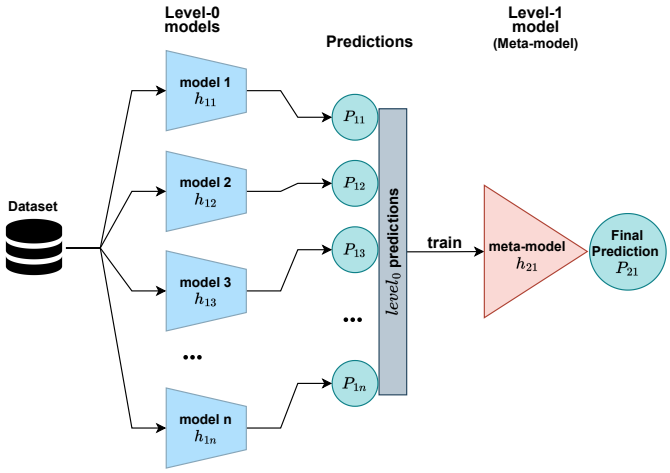


Fig. 9: Stacking ensemble learning architecture.

the sequential order differences. By combining these outputs and training our meta-model, meMIA provides a comprehensive understanding of dataset characteristics for accurate membership inference, subsequently increasing the potency of the attack.

The base neural network model comprises of a 4-layer, fully connected architecture with 512, 256, 128, and 64 layer dimensions. Complementing this, the base RNN-LSTM model is designed as a 3 multi-layer LSTM, with layer sizes of 256, 128, and 50 as depicted in Figure 3. Additionally, the meta-model processes input embeddings with dimensions [64+50+$z$], where $z$ is the prediction of the shadow model that is constructed as a 4-layer, fully connected neural network, with layer sizes set to 512, 256, 128, and 2 respectively.

We chose a mini-batch size 32 and used cross-entropy as the loss function for the base models. For the meta-model, we used binary cross-entropy. Moreover, we use the Adam optimizer with a learning rate of $1e-5$, set ReLU as the activation function across the three networks, and jointly train base and meta models for 80 epochs.

TABLE IV: Comparative analysis of accuracy: Membership inference attack models vs. meMIA.

| Dataset | NSH Acc. [34] | MIA Acc. [45] | seqMIA Acc. [29] | meMIA Acc. |
|---|---|---|---|---|
| FMNIST | 0.568 | 0.577 | 0.579 | **0.598** |
| UTKFace | 0.588 | 0.589 | 0.594 | **0.624** |
| STL-10 | 0.785 | 0.808 | 0.820 | **0.853** |
| CIFAR-10 | 0.776 | 0.791 | 0.792 | **0.827** |
| CIFAR-100 | 0.886 | 0.906 | 0.909 | **0.946** |
| Location | 0.594 | 0.750 | 0.759 | **0.783** |
| Purchase-100 | 0.734 | 0.800 | 0.803 | **0.822** |

### A. *Scalability and complexity of meMIA*

In this section, we explore the impact of the added layer of sophistication in terms of $level_0$ models and discuss it in terms of time efficiency and its scalability towards sophisticated datasets and then compare the performance with the baseline MIA attack.

First, we examine the number of parameters used in the training by each state-of-the-art (SOTA) attack method. MIA's strategy involves training n-inference models that collectively act as an attack model, each with a single layer of 128 neurons, where n is the number of classes for a given dataset. Therefore, the total number of parameters, $ps$, that MIA trains are $n[128\mathcal{X}+386]$, where $n$ is the number of classes and $\mathcal{X}$ is the prediction vector of size equal to the number of classes. Thus, replacing $\mathcal{X}$ with $n$, we get $ps = 128n^2 + 386$. Hence, the complexity of the MIA depends upon the number of classes $n$, leading to a squared asymptotic growth with respect to the number of classes, i.e., $\mathcal{O}(n^2)$.

Additionally, it's clear that the time complexity of the model depends upon the number of computing units (neurons). Thus, training time is proportional to the total number of computing

units. However, meMIA is agnostic to the class parameter $n$ in terms of time complexity because meMIA's architecture does not change regardless of the dataset type. Consequently, meMIA's time complexity remains considerably close to MIA, considering a slight variance in execution time depending on the complexity of the dataset itself, not the number of classes. Moreover, as shown in Figure 10, we observe around a 20 seconds higher training time compared to baseline MIA, as the additional execution overhead comes from the LSTM layers, which are particularly resource-intensive due to their sequential data processing and internal complexity. Comparatively, meMIA, with its added layer of sophistication, holds a clear advantage with comparable computational overhead, considering the overall inference performance gain across the datasets.

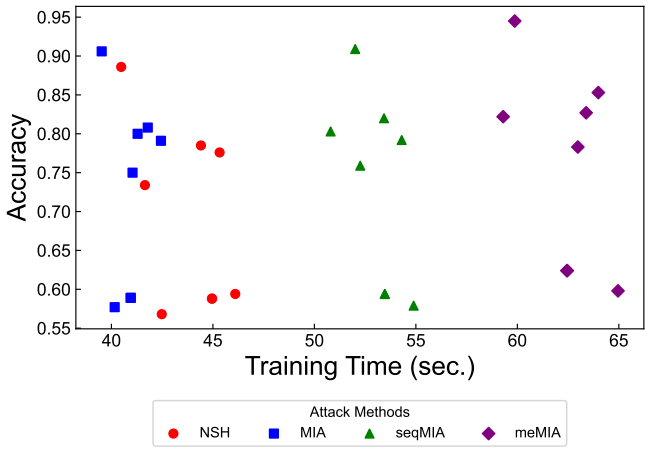


Fig. 10: Comparison of training time and accuracy across different attack methods.

### B. Ablation study of meMIA

To elaborate on and comprehend the individual and combined contributions to meMIA's inference attack performance, we conducted an ablation study across the benchmark datasets. Moreover, to measure the effect of each component, we selectively disabled the components and ran experiments to observe their impact on inference gain.

**Impact of the NN component:** When we drop the LSTM component, we observe a notable decrease in accuracy across all the datasets, as illustrated in Table V. This suggests that NN is only capable of learning static patterns. Additionally, NN lacks sequential learning capabilities, which the LSTM provides. The absence of LSTM led to an obvious drop in accuracy, emphasizing the importance of treating prediction vectors as sequential features compared to solely learning from the difference in the prediction vector distribution.

**Impact of the LSTM component:** On the other hand, if we disable the NN component and retain the LSTM, we notice a similar trend of reduced accuracy. These results evidently signify that the LSTM component alone, despite its proficiency in learning the order dependencies, is not able to fully capture the intricate interaction of confidence scores. Moreover, the comparison shows that the NN components generally outperform the LSTM in capturing the distribution differences at the output of the shadow model, suggesting that NN contributes more to the combined attack.

**Combined Impact in meMIA:** When the two components are ensembled, meMIA consistently outperforms the individual components. With the full meMIA model, the accuracy improvements are quite significant, as shown in Table IV. Overall, the experiments have shown that meMIA's ensemble approach not only preserved the individual strengths of NN and LSTM but also enhanced the effectiveness of inference attacks when combined together.

TABLE V: Comparative accuracy metrics for meMIA with isolated and integrated NN and LSTM components.

| Dataset | $model_0$ NN Acc. | $model_1$ LSTM Acc. | meMIA Acc. |
|---|---|---|---|
| FMNIST | 57.64 | 51.07 | **59.8** |
| UTKFace | 62.04 | 58.07 | **62.4** |
| STL-10 | 79.98 | 77.79 | **85.3** |
| CIFAR-10 | 78.9 | 76.62 | **82.7** |
| CIFAR-100 | 91.36 | 88.55 | **94.6** |
| Location | 68.75 | 69.78 | **78.3** |
| Purchase-100 | 81.22 | 52.73 | **82.2** |

## V. EXPERIMENTAL SETUP

The training datasets for the target and shadow models are chosen at random and equally divided, ensuring that there are no overlaps between the two sets (see section III). We used randomly shuffled disjoint target and shadow training samples in our experiments to evaluate meMIA. In other words, similar to [45], we did not use any sample in the training of our attack model, which was used to train the target model, thereby ensuring that the attacker has little knowledge of the actual training data. More precisely, the training samples of the target and attack models are mutually exclusive, increasing the odds of uncertainty setting up the most challenging attack setting, leaving the adversary with a baseline accuracy of 50% equivalent to random guessing.

In this paper, we use standard metrics such as precision and recall, where precision shows how accurately the attack model predicts a fraction of samples being members and recall measures the coverage. Moreover, We present both class-specific and average measurements to compare meMIA's performance with state-of-the-art attacks, using MIA [45] as our *baseline* attack model. Note, when we mention MIA in our experiments, we refer to [45].

The experiments on all datasets were run 5 times, reporting mean values, against our target models trained on a local PC equipped with one NVIDIA T600 GPU, 16 GB of memory and an Intel E-cores CPU server with Ubuntu 22.04 operating system.

## VI. RESULTS AND DISCUSSION

In this section we first discuss techniques we used to compare meMIA's inference performance with SOTA MIA inference schemes. Next, we discuss our results.

### A. Techniques used for comparison

We used several methods to compare the performance of meMIA attack model. As mention previously we chose SOTA inference schemes MIA [45], seqMIA [28] and NSH [34] to show the superior inference capabilities of meMIA under the extremely difficult attack scenario (the Black box setting) over diverse range of datasets using the following performance metrics:

**Cumulative Fraction of Classes (CDF) for precision and recall:** To calculate the cumulative distribution function (CDF) of accuracy values across different classes given a particular dataset, we do the following:

- First we sort these precision values in ascending order
- Next, for each of the sorted precision value, we calculate the cumulative probability which is the sum of the probabilities of all precision less than or equal to that value. for example if we have 5 classes with precision values of 0.70, 0.82, 0.65, 0.90, and 0.75, we consider each having an equal probability of $1/5$ as shown below:
  - $P(X \leq 0.65) = 1/5$
  - $P(X \leq 0.70) = 1/5$
  - $P(X \leq 0.75) = 1/5$
  - $P(X \leq 0.82) = 1/5$
  - $P(X \leq 0.90) = 1/5$

Besides, we also determine CDF of recall using the same procedure. Moreover, The CDF provides us a clear understanding of how precision and recall values are distributed across the classes, enabling us to identify whether most classes are easy or difficult for the attack model to infer. Henceforth we can use CDF to compare multiple MIA attack methods and observe their overall performance across all classes.

**Accuracy and AUC:** As mentioned previously, we use precision and recall, which are standard metrics used across MIA (Membership Inference Attack) literature. However, we also use accuracy and AUC (Area Under the ROC Curve) to provide a comprehensive evaluation of meMIA.

### B. Results

The accuracy of target models for different datasets is shown in Table III. For some datasets with the same hyperparameter set, the accuracy reaches as low as 27%. For less complex datasets like FMNIST, the attack inference of all state-of-the-art attacks is moderate, as depicted in Table IV. Specifically, for FMNIST, our attack model achieves an attack accuracy that is **3.63%** higher than the baseline MIA attack. The improvement is even more significant for datasets like UTKFace and STL-10, reaching a **6%** increase. Additionally, with complex datasets like CIFAR-100, our attack model stands out, showing a **94.6%** accuracy rate.

In Figure 4(a), the median precision (indicated by the line within each box) of our attack method is consistently higher across all three dataset size variations compared to MIA [45]. This suggests that meMIA typically achieves superior inference performance with higher median precision and exhibits less variability. Similarly, when we compare the CDF curves for both attack methods as shown in Figure 4(b), meMIA outperforms MIA, indicating that meMIA not only correctly identifies more true members of the dataset (higher recall) but also makes these identifications with fewer false positives (higher precision).

Figure 5(a) shows precision and recall for membership inference against CNN-based target neural networks trained on STL-10 dataset of size 6,500, illustrating a comparison between MIA and meMIA methods per class performance. We observe fluctuations in performance due to varying number of samples per class and feature distribution. Furthermore, we also infer that meMIA does reasonably well as both precision and accuracy are mostly above **70%** compared to the baseline MIA attack. Additionally, Figure 5(b) shows the cumulative fraction of classes at different accuracy levels for both the baseline MIA and meMIA on a CNN trained on the STL-10 dataset. The effective CDF distribution of meMIA indicates it can identify actual members with fewer errors across more classes, demonstrating its potency and efficiency in membership inference attacks.

Figure 6 presents the precision and recall metrics as cumulative distribution function (CDF) graphs for the CIFAR-100 dataset. Figure 6(a) compares baseline MIA and Figure 6(b) compares seqMIA [29] with our attack method, where meMIA consistently outperforms them, indicating our model is more effective in precision and recall. Additionally, it shows a strong ability to conduct membership inference attacks on the CIFAR-100 dataset. Similarly, when we perform experiments on the Purchase-100 dataset, we can observe the same trend where our model outperforms MIA and seqMIA as shown in Figure 7(a) and Figure 7(b) .

Figure 8 shows the performance of membership inference attacks on a CNN trained with the FMNIST (Fashion-MNIST) dataset, comparing the CDFs of MIA and seqMIA with meMIA. Furthermore, it can be observed from the graphs that the precision of our attack surpasses MIA at an accuracy threshold of approximately 52%, suggesting that our method can correctly identify true positives more efficiently than MIA beyond this accuracy threshold as shown in Figure 8(a). Similarly, our attack achieves greater precision than seqMIA, around an accuracy of 57%, with both methods showing comparable precision up to this accuracy level as depicted in Figure 8(b). Moreover, we note that all attack models exhibit limited success, which can be attributed to the 90% high test accuracy of the CNN-based target model for FMNIST data. In other words, target models generalized the data very well, leaving a very small window for the success of the attacks.

*1) Attack performance Against Dataset Complexity:* In this paper, we used seven benchmark datasets and three attack methods in a black box setting. We performed extensive experiments to validate the efficacy of the proposed membership inference attack model. To maintain the same level of features for the image-based dataset, we resized all the samples to 32x32 pixels, as mentioned before (see III-A) and observed the following.

- As shown in Figure 11, the complexity of the dataset has a tremendous effect on the inference performance of the attack methods. FMNIST is the simplest dataset, followed by UTKFace, which contains colour images of human faces (4 classes based on races have been used),

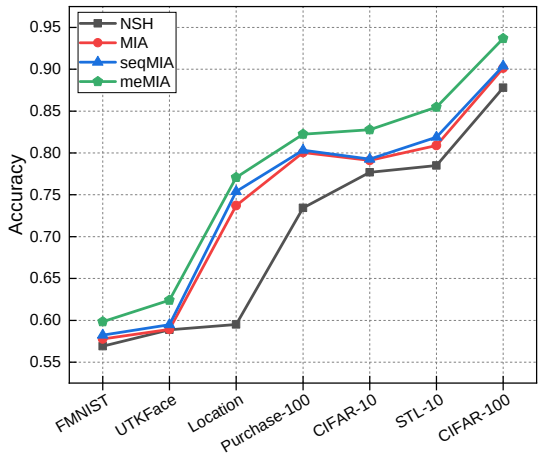


Fig. 11: Comparative performance of four attack models across seven benchmark datasets with varying levels of data complexity.

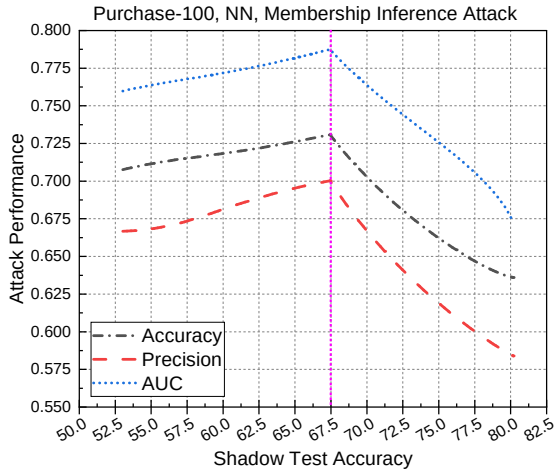


Fig. 12: Membership inference Attack performance (using accuracy, precision and AUC metrics) against target model for Purchase-100 dataset with 0.6602 testing accuracy and changing the shadow model test accuracy, illustrating the best performance with lowest contrast.

and so on until CIFAR-100, the most complex dataset. More precisely, as the complexity of the dataset increases, membership inference attack models gain better attack performance.

- We observe that our proposed attack model consistently outperforms all three attack methods while following the trend of acquiring more potency as the dataset gets complex, showcasing the robustness and versatility of meMIA model. In particular, we observe a significant increase in performance on the CIFAR-10, STL-10 and CIFAR-100 datasets compared to the other models, indicating that our model generalizes better as the dataset becomes complex.

*2) Relation Among Testing Accuracy of Target and Shadow Models:* To analyze how attack inference performance responds to varying the test accuracy of the shadow model (by overfitting or underfitting) while keeping the target model's testing accuracy fixed at 66.02%. We observe that the attack model performs best when the shadow model has a slightly lower accuracy than the target model, as shown in the Figure 12. The peak performance occurs when the shadow model's accuracy is approximately 62%-65%, which is the sweet spot for the attack model on the Purchase-100 dataset. More precisely, the trend in the Figure shows that a high contrast between the shadow and target model's testing accuracies will lead to degraded membership inference attack performance, confirming the experiments demonstrated by [41].

## VII. CONCLUSION

In this paper, we introduced meMIA, a novel membership inference attack model. The key insight while devising our attack model involved leveraging multi-level information using stacked ensemble machine learning. We implemented three different threat models, i.e., NSH, MIA and seqMIA, within a black-box attack framework to assess the relative potency of meMIA. For target model architectures, we considered a simple CNN for image-based datasets and a simple NN for location and purchase datasets. We conducted extensive experiments to validate the effectiveness of our model using seven benchmark datasets. The results show that meMIA outperforms state-of-the-art MI attacks, achieving accuracy as high as 94% and near-perfect recall. Among other things, our analysis shows a strong correlation between the test accuracies of the target and shadow models and how that affects the performance of the attack model. For example, the attack model performs better if the difference in accuracy between the target and shadow models approaches zero. We further affirm that as dataset complexity increases, the attack model's performance improves.

As part of future work, it would be interesting to delve deeper and experiment with more advanced machine learning architectures, e.g., designing transformer-based "attention is all you need" attack models. Additionally, to further improve our attack model, we intend to explore representation learning to learn two distinct latent feature spaces for member and non-member data samples. Moreover, to make meMIA resilient against potential defensive mechanisms, exploring the use of gradient masking techniques in the training of shadow models could further enhance meMIA's inference performance.

## REFERENCES

[1] *[1702.08423] Age Progression/Regression by Conditional Adversarial Autoencoder*. URL: https://arxiv.org/abs/1702.08423 (visited on 12/19/2023).

[2] *[2106.11384] Membership Inference on Word Embedding and Beyond*. URL: https://arxiv.org/abs/2106.11384 (visited on 07/06/2024).

[3] *[2107.06018] This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces*. URL: https://arxiv.org/abs/2107.06018 (visited on 07/04/2024).

[4] *[2108.03506] Membership Inference Attacks on Lottery Ticket Networks*. URL: https://arxiv.org/abs/2108.03506 (visited on 07/06/2024).

[5] *[PDF] Learning Multiple Layers of Features from Tiny Images — Semantic Scholar*. URL: https://www.semanticscholar.org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086 (visited on 12/19/2023).

[6] Martín Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. arXiv:1607.00133 [cs, stat]. Oct. 2016, pp. 308–318. DOI: 10.1145/2976749.2978318. URL: http://arxiv.org/abs/1607.00133 (visited on 04/24/2024).

[7] Hassan Ali et al. "Membership Inference Attacks on DNNs using Adversarial Perturbations". In: *arXiv.org* (2023). DOI: 10.48550/arxiv.2307.05193.

[8] Mauro Barni et al. "Privacy-Preserving ECG Classification With Branching Programs and Neural Networks". In: *IEEE Transactions on Information Forensics and Security* 6.2 (June 2011). Conference Name: IEEE Transactions on Information Forensics and Security, pp. 452–468. ISSN: 1556-6021. DOI: 10.1109/TIFS.2011.2108650. URL: https://ieeexplore.ieee.org/document/5702365 (visited on 04/24/2024).

[9] Hannah Brown et al. *What Does it Mean for a Language Model to Preserve Privacy?* arXiv:2202.05520 [cs, stat]. Feb. 2022. DOI: 10.48550/arXiv.2202.05520. URL: http://arxiv.org/abs/2202.05520 (visited on 04/24/2024).

[10] Kamalika Chaudhuri and Claire Monteleoni. "Privacy-preserving logistic regression". In: *Advances in Neural Information Processing Systems*. Vol. 21. Curran Associates, Inc., 2008. URL: https://papers.nips.cc/paper_files/paper/2008/hash/8065d07da4a77621450aa84fee5656d9-Abstract.html (visited on 04/24/2024).

[11] Dingfan Chen et al. "GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models". In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. arXiv:1909.03935 [cs]. Oct. 2020, pp. 343–362. DOI: 10.1145/3372297.3417238. URL: http://arxiv.org/abs/1909.03935 (visited on 12/19/2023).

[12] Min Chen et al. "When Machine Unlearning Jeopardizes Privacy". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. arXiv:2005.02205 [cs, stat]. Nov. 2021, pp. 896–911. DOI: 10.1145/3460120.3484756. URL: http://arxiv.org/abs/2005.02205 (visited on 12/19/2023).

[13] Adam Coates, Andrew Ng, and Honglak Lee. "An Analysis of Single-Layer Networks in Unsupervised Feature Learning". en. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. JMLR Workshop and Conference Proceedings, June 2011, pp. 215–223. URL: https://proceedings.mlr.press/v15/coates11a.html (visited on 12/19/2023).

[14] Mauro Conti, Jiaxin Li, and Stjepan Picek. *On the Vulnerability of Data Points under Multiple Membership Inference Attacks and Target Models*. arXiv:2210.16258 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2210.16258. URL: http://arxiv.org/abs/2210.16258 (visited on 07/06/2024).

[15] Emiliano De Cristofaro. *An Overview of Privacy in Machine Learning*. arXiv:2005.08679 [cs, stat]. May 2020. DOI: 10.48550/arXiv.2005.08679. URL: http://arxiv.org/abs/2005.08679 (visited on 12/19/2023).

[16] Cynthia Dwork. "Differential Privacy". en. In: *Encyclopedia of Cryptography and Security*. Ed. by Henk C. A. van Tilborg and Sushil Jajodia. Boston, MA: Springer US, 2011, pp. 338–340. ISBN: 978-1-4419-5906-5. DOI: 10.1007/978-1-4419-5906-5_752. URL: https://doi.org/10.1007/978-1-4419-5906-5_752 (visited on 04/24/2024).

[17] James Franklin. "The elements of statistical learning: data mining, inference and prediction". en. In: *The Mathematical Intelligencer* 27.2 (Mar. 2005), pp. 83–85. ISSN: 0343-6993. DOI: 10.1007/BF02985802. URL: https://doi.org/10.1007/BF02985802 (visited on 12/12/2023).

[18] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1322–1333. ISBN: 978-1-4503-3832-5. DOI: 10.1145/2810103.2813677. URL: https://dl.acm.org/doi/10.1145/2810103.2813677 (visited on 12/05/2023).

[19] Xinlei He et al. *Node-Level Membership Inference Attacks Against Graph Neural Networks*. arXiv:2102.05429 [cs]. Feb. 2021. DOI: 10.48550/arXiv.2102.05429. URL: http://arxiv.org/abs/2102.05429 (visited on 07/06/2024).

[20] Xinlei He et al. *Stealing Links from Graph Neural Networks*. arXiv:2005.02131 [cs]. Oct. 2020. DOI: 10.48550/arXiv.2005.02131. URL: http://arxiv.org/abs/2005.02131 (visited on 12/06/2023).

[21] Nils Homer et al. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays". eng. In: *PLoS genetics* 4.8 (Aug. 2008), e1000167. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000167.

[22] Dazhi Hu. *Membership Inference Attacks Against Latent Factor Model*. arXiv:2301.03596 [cs]. Dec. 2022. DOI: 10.48550/arXiv.2301.03596. URL: http://arxiv.org/abs/2301.03596 (visited on 07/04/2024).

[23] Hailong Hu and Jun Pang. *Membership Inference of Diffusion Models*. arXiv:2301.09956 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2301.09956. URL: http://arxiv.org/abs/2301.09956 (visited on 07/04/2024).

[24] Klas Leino and Matt Fredrikson. *Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference*. arXiv:1906.11798 [cs, stat]. June 2020. DOI: 10.48550/arXiv.1906.11798. URL: http://arxiv.org/abs/1906.11798 (visited on 12/19/2023).

[25] Zheng Li and Yang Zhang. *Membership Leakage in Label-Only Exposures*. arXiv:2007.15528 [cs, stat]. Sept. 2021. DOI: 10.48550/arXiv.2007.15528. URL: http://arxiv.org/abs/2007.15528 (visited on 12/19/2023).

[26] Yehuda Lindell and Benny Pinkas. "Privacy Preserving Data Mining". en. In: *Advances in Cryptology — CRYPTO 2000*. Ed. by Mihir Bellare. Berlin, Heidelberg: Springer, 2000, pp. 36–54. ISBN: 978-3-540-44598-2. DOI: 10.1007/3-540-44598-6_3.

[27] Junlin Liu et al. "Similarity-based Label Inference Attack against Training and Inference of Split Learning". In: *IEEE Transactions on Information Forensics and Security* 19 (2024). arXiv:2203.05222 [cs], pp. 2881–2895. ISSN: 1556-6013, 1556-6021. DOI: 10.1109/TIFS.2024.3356821. URL: http://arxiv.org/abs/2203.05222 (visited on 07/06/2024).

[28] Yaru Liu and Junyi Xin. "SeqMIA: Membership Inference Attacks Against Machine Learning Classifiers Using Sequential Information". en. In: *2022 4th International Conference on Applied Machine Learning (ICAML)*. Changsha, China: IEEE, July 2022, pp. 31–35. ISBN: 978-1-66546-265-5. DOI: 10.1109/ICAML57167.2022.00014. URL: https://ieeexplore.ieee.org/document/10056445/ (visited on 09/30/2023).

[29] Yaru Liu and Junyi Xin. "SeqMIA: Membership Inference Attacks Against Machine Learning Classifiers Using Sequential Information". In: *2022 4th International Conference on Applied Machine Learning (ICAML)* (2022). DOI: 10.1109/icaml57167.2022.00014.

[30] Yugeng Liu et al. *ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models*. arXiv:2102.02551 [cs, stat]. Oct. 2021. DOI: 10.48550/arXiv.2102.02551. URL: http://arxiv.org/abs/2102.02551 (visited on 12/06/2023).

[31] Yunhui Long et al. *Understanding Membership Inferences on Well-Generalized Learning Models*. arXiv:1802.04889 [cs, stat]. Feb. 2018. DOI: 10.48550/arXiv.1802.04889. URL: http://arxiv.org/abs/1802.04889 (visited on 12/07/2023).

[32] Luca Melis et al. *Exploiting Unintended Feature Leakage in Collaborative Learning*. arXiv:1805.04049 [cs]. Nov. 2018. DOI: 10.48550/arXiv.1805.04049. URL: http://arxiv.org/abs/1805.04049 (visited on 12/06/2023).

[33] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning". In: *2019 IEEE Symposium on Security and Privacy (SP)*. arXiv:1812.00910 [cs, stat]. May 2019, pp. 739–753. DOI: 10.1109/SP.2019.00065. URL: http://arxiv.org/abs/1812.00910 (visited on 12/07/2023).

[34] Milad Nasr, Reza Shokri, and Amir Houmansadr. *Machine Learning with Membership Privacy using Adversarial Regularization*. arXiv:1807.05852 [cs, stat]. July 2018. DOI: 10.48550/arXiv.1807.05852. URL: http://arxiv.org/abs/1807.05852 (visited on 12/07/2023).

[35] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. *Knockoff Nets: Stealing Functionality of Black-Box Models*. arXiv:1812.02766 [cs]. Dec. 2018. DOI: 10.48550/arXiv.1812.02766. URL: http://arxiv.org/abs/1812.02766 (visited on 12/19/2023).

[36] Nicolas Papernot et al. *Towards the Science of Security and Privacy in Machine Learning*. arXiv:1611.03814 [cs]. Nov. 2016. DOI: 10.48550/arXiv.1611.03814. URL: http://arxiv.org/abs/1611.03814 (visited on 12/06/2023).

[37] Sankita Patel, Viren Patel, and Devesh Jinwala. "Privacy Preserving Distributed K-Means Clustering in Malicious Model Using Zero Knowledge Proof". en. In: *Distributed Computing and Internet Technology*. Ed. by Chittaranjan Hota and Pradip K. Srimani. Berlin, Heidelberg: Springer, 2013, pp. 420–431. ISBN: 978-3-642-36071-8. DOI: 10.1007/978-3-642-36071-8_33.

[38] Vinayakumar Ravi, Vasundhara Acharya, and Mamoun Alazab. "A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images". en. In: *Cluster Computing* 26.2 (Apr. 2023), pp. 1181–1203. ISSN: 1573-7543. DOI: 10.1007/s10586-022-03664-6. URL: https://doi.org/10.1007/s10586-022-03664-6 (visited on 12/11/2023).

[39] Shahbaz Rezaei and Xin Liu. *On the Discredibility of Membership Inference Attacks*. arXiv:2212.02701 [cs]. Apr. 2023. DOI: 10.48550/arXiv.2212.02701. URL: http://arxiv.org/abs/2212.02701 (visited on 07/06/2024).

[40] Benjamin I. P. Rubinstein et al. *Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning*. arXiv:0911.5708 [cs]. Nov. 2009. DOI: 10.48550/arXiv.0911.5708. URL: http://arxiv.org/abs/0911.5708 (visited on 04/24/2024).

[41] Ahmed Salem et al. *ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models*. arXiv:1806.01246 [cs]. Dec. 2018. DOI: 10.48550/arXiv.1806.01246. URL: http://arxiv.org/abs/1806.01246 (visited on 12/06/2023).

[42] Ahmed Salem et al. *Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning*. arXiv:1904.01067 [cs, stat]. Nov. 2019. DOI: 10.48550/arXiv.1904.01067. URL: http://arxiv.org/abs/1904.01067 (visited on 12/06/2023).

[43] Sriram Sankararaman et al. "Genomic privacy and limits of individual detection in a pool". en. In: *Nature Genetics* 41.9 (Sept. 2009). Number: 9 Publisher: Nature Publishing Group, pp. 965–967. ISSN: 1546-1718. DOI: 10.1038/ng.436. URL: https://www.nature.com/articles/ng.436 (visited on 12/07/2023).

[44] Reza Shokri and Vitaly Shmatikov. "Privacy-Preserving Deep Learning". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1310–1321. ISBN: 978-1-4503-3832-5. DOI: 10.1145/2810103.2813687. URL: https://doi.org/10.1145/2810103.2813687 (visited on 04/24/2024).

[45] Reza Shokri et al. *Membership Inference Attacks against Machine Learning Models*. arXiv:1610.05820 [cs, stat]. Mar. 2017. DOI: 10.48550/arXiv.1610.05820.

URL: http://arxiv.org/abs/1610.05820 (visited on 12/06/2023).

[46] Congzheng Song and Vitaly Shmatikov. *Overlearning Reveals Sensitive Attributes*. arXiv:1905.11742 [cs, stat]. Feb. 2020. DOI: 10.48550/arXiv.1905.11742. URL: http://arxiv.org/abs/1905.11742 (visited on 12/06/2023).

[47] Anshuman Suri et al. *Subject Membership Inference Attacks in Federated Learning*. arXiv:2206.03317 [cs]. June 2023. DOI: 10.48550/arXiv.2206.03317. URL: http://arxiv.org/abs/2206.03317 (visited on 07/06/2024).

[48] Florian Tramèr et al. *Stealing Machine Learning Models via Prediction APIs*. arXiv:1609.02943 [cs, stat]. Oct. 2016. DOI: 10.48550/arXiv.1609.02943. URL: http://arxiv.org/abs/1609.02943 (visited on 12/06/2023).

[49] Yu Wang et al. *Membership Inference Attacks on Knowledge Graphs*. arXiv:2104.08273 [cs]. Feb. 2022. DOI: 10.48550/arXiv.2104.08273. URL: http://arxiv.org/abs/2104.08273 (visited on 07/06/2024).

[50] David H. Wolpert. "Stacked generalization". In: *Neural Networks* 5.2 (Jan. 1992), pp. 241–259. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(05)80023-1. URL: https://www.sciencedirect.com/science/article/pii/S0893608005800231 (visited on 12/28/2023).

[51] Yixin Wu et al. *Membership Inference Attacks Against Text-to-image Generation Models*. arXiv:2210.00968 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2210.00968. URL: http://arxiv.org/abs/2210.00968 (visited on 07/06/2024).

[52] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. arXiv:1708.07747 [cs, stat]. Sept. 2017. DOI: 10.48550/arXiv.1708.07747. URL: http://arxiv.org/abs/1708.07747 (visited on 12/11/2023).

[53] Pengtao Xie et al. *Crypto-Nets: Neural Networks over Encrypted Data*. arXiv:1412.6181 [cs]. Dec. 2014. DOI: 10.48550/arXiv.1412.6181. URL: http://arxiv.org/abs/1412.6181 (visited on 04/24/2024).

[54] Mingxue Xu and Xiang-Yang Li. *Data Origin Inference in Machine Learning*. arXiv:2211.13416 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2211.13416. URL: http://arxiv.org/abs/2211.13416 (visited on 07/06/2024).

[55] Dingqi Yang, Daqing Zhang, and Bingqing Qu. "Participatory Cultural Mapping Based on Collective Behavior Data in Location-Based Social Networks". en. In: *ACM Transactions on Intelligent Systems and Technology* 7.3 (Apr. 2016), pp. 1–23. ISSN: 2157-6904, 2157-6912. DOI: 10.1145/2814575. URL: https://dl.acm.org/doi/10.1145/2814575 (visited on 12/19/2023).

[56] Yunhao Yang, Parham Gohari, and Ufuk Topcu. "On The Vulnerability of Recurrent Neural Networks to Membership Inference Attacks". en. In: (Oct. 2021). URL: https://openreview.net/forum?id=sBHVNmCt3t (visited on 07/06/2024).

[57] Samuel Yeom et al. *Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting*. arXiv:1709.01604 [cs, stat]. May 2018. DOI: 10.48550/arXiv.1709.01604. URL: http://arxiv.org/abs/1709.01604 (visited on 12/07/2023).

[58] Cha Zhang and Yunqian Ma, eds. *Ensemble Machine Learning: Methods and Applications*. en. New York, NY: Springer, 2012. ISBN: 978-1-4419-9325-0 978-1-4419-9326-7. DOI: 10.1007/978-1-4419-9326-7. URL: https://link.springer.com/10.1007/978-1-4419-9326-7 (visited on 12/11/2023).

[59] Jun Zhang et al. "Functional mechanism: regression analysis under differential privacy". In: *Proceedings of the VLDB Endowment* 5.11 (July 2012), pp. 1364–1375. ISSN: 2150-8097. DOI: 10.14778/2350229.2350253. URL: https://doi.org/10.14778/2350229.2350253 (visited on 04/24/2024).

[60] Peng Zhang et al. "Privacy Preserving Naive Bayes Classification". en. In: *Advanced Data Mining and Applications*. Ed. by Xue Li, Shuliang Wang, and Zhao Yang Dong. Berlin, Heidelberg: Springer, 2005, pp. 744–752. ISBN: 978-3-540-31877-4. DOI: 10.1007/11527503_88.

[61] Yuheng Zhang et al. *The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks*. arXiv:1911.07135 [cs, stat]. Apr. 2020. DOI: 10.48550/arXiv.1911.07135. URL: http://arxiv.org/abs/1911.07135 (visited on 12/06/2023).

[62] Junhao Zhou et al. *Property Inference Attacks Against GANs*. arXiv:2111.07608 [cs, stat]. Nov. 2021. DOI: 10.48550/arXiv.2111.07608. URL: http://arxiv.org/abs/2111.07608 (visited on 07/06/2024).

**Najeeb Ullah** is a Ph.D. student at the Department of Electrical and Computer Engineering at the National University of Singapore. He received his B.Sc. degree in Computer Science from FAST-NUCES, Peshawar, Pakistan, in 2017 and M.Sc. degree in Computer Science from FAST-NUCES, Peshawar, Pakistan, in 2019. His research interests include cybersecurity and quantum machine learning.

**Muhammad Naveed Aman** (S'12-M'17-SM'23) is an Assistant Professor in the University of Nebraska-Lincoln. He received the B.Sc. degree in Computer Systems Engineering from KPK UET, Peshawar, Pakistan, M.Sc. degree in Computer Engineering from the Center for Advanced Studies in Engineering, Islamabad, Pakistan, M.Engg. degree in Industrial and Management Engineering and Ph.D. in Electrical Engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA in 2006, 2008, and 2012, respectively. His research interests include IoT and network security, hardware systems security and privacy, wireless and mobile networks and stochastic modelling.

**Biplab Sikdar** received the B.Tech. degree in electronics and communication engineering from North Eastern Hill University, Shillong, India, in 1996, the M.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1998, and the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2001. He is currently a Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include wireless network, and security for IoT and cyber physical systems.