**Literature Review:**

Machine learning has played crucial roles in enhancing great business impacts on different organizations in the world. This advanced technology processes images, text, audio, and video. Membership inference attacks challenge the privacy of machine learning models by determining if a specific data record was part of the model's training dataset. To figure out if a specific piece of data was used to train a machine learning model, create another model that can spot differences in how the original model responds to data it has seen before versus data it hasn't seen. This new model helps us identify whether the original model was trained on a particular input or not (Liu et al., 2023). Performing an attack like membership inference needs two models—a target model and an inference or attack model—that check the differences in response in the target model in the same input data that is trained on versus the data that is unseen or untrained on the target model (Shokriet al., 2017a).

Based on classification models that outsource from commercial machine learning platforms such as Google and Amazon, realistic data, like hospital discharge data, is vulnerable to membership inference attacks. Investigate the key roles in this type of attack and explore mitigation from this attack (Shokri et al., 2017b). For example, knowing that a patient's clinical record was used to train a model associated with a specific disease could reveal that the patient has that disease. The success of these attacks is dependent on how it works with unseen data—handle a variety of inputs and provide reliable outputs. This success depends on the diversity of training datasets; then it works well with unknown data that has not been trained. The model can be overfitted when it takes too much detail and noise from training data, which degrades performance while performing with non-trained or unseen data. Another problem is non-representative data, which basically occurs when training data is narrow or biased; it cannot effectively handle diverse inputs, making it more prone to errors and vulnerabilities.

Models that learn too much detail from their training data or are trained on biased data are easier to attack. Overfitting means a model works well with the data it was trained on but fails with new data. The models that overfit, or those trained on non-representative data, are more vulnerable to these attacks. Overfitting is when a model performs well on training data but poorly on new data. Overfitted data show different behaviors, which are more susceptible to MIA attacks. The number of output classes in a model also influences its vulnerabilities. If the models have many classes, they leak more information because they can extract specific features from the data, which helps them accurately classify the inputs. A study showed that although the number of training records per class is the same, the attacks are more effective against CIFAR-100 than CIFAR-10 because they have more classes and are overfitted (Shokri et al., 2017c). The amount and diversity of training data play a vital role regarding attack success. Generally, more training data for a class makes it harder for attackers to infer membership because it shows better generalization and variability in each class. For this reason, the lack of randomness of training data minimizes the chance of a membership inference attack.

Moreover, different machine learning models leak different amounts of information from their training data because of their unique designs. How well a model generalizes also affects how easily membership inference attacks can succeed. It is a fact that models that are more overfitted disclose more information, but this is only true for some types of models. A good illustration is that an overfitted Amazon model leaks more information than a less overfitted Amazon model. However, both Amazon models leak less than a Google model, even though the Google model is less overfitted and has better predictive power. As a result, overfitting isn’t the only reason for information leakage; the model’s structure and type also played a pivotal role.

**Types of membership inference attacks:** There are some common types of membership inference attacks in federated learning that determine a specific data point was part of the training dataset in an FL model.

Black-box Attacks:The attackers do this attack process by querying the model with specific data points and observing the model’s output. After getting the confidence score and probabilities returned by the model output, the advisory can infer whether a particular data point was part of the training set. Generally, models are more confident when predicting training data compared to non-training data.

White-box Attacks: The attackers try to gain access to the model’s internal parameters and gradients, allowing them to make more precise membership inferences. In fact, the attacker has detailed access to the model's inner workings. The advisory knows the model structure (layer), model parameters (weight and biases), and training details (used algorithms and number of training rounds).

Shadow Model Attacks: This attack method mimics the target model behavior, but it works with a known training dataset. The attack model is then trained with labeled inputs and takes output from the shadow model. Various methods are used to generate training data for the shadow models, including black box access to the target model, population statistics, and noisy dataset versions.

Label-only Attacks: The attacker only sees the predicted labels, not the confidence scores. Even with just the labels, the attacker can infer membership by observing how the model behaves with different inputs.

Subject Membership Inference Attacks: The attackers try to determine if a specific subject (like a patient in a medical data set) is part of the training data. The attackers use detailed knowledge of the model and training process to make this determination, often leveraging access to the model after each training round.

**Key Studies and Findings of Existing Literature:** We investigate some papers to find out the current literature and findings of the studies and try to understand the shortcomings of the existing literature. Basically, we discovered some types of MIA attacks and analyzed how they work. Among them, researchers have developed a new technique to train attack models called shadow training. We create several shadow models that mimic the target model's behavior (Shokri et al., 2017a). For these shadow models, we know the training datasets, so we can accurately determine which data points are part of these datasets. This method has distinguished members from non-members of a training dataset of a target model. The attack models are trained using proxy targets, where the training data is known, allowing for effective supervised training.

A diagram of a model

AI-generated content may be incorrect.

Fig 1: Shadow training architecture.

First, several shadow models are trained using their own datasets to mimic the target model's behavior. Then, a meta-model is trained using the outputs from these shadow models and the known labels from their datasets. This meta-model helps determine if data belongs to the target model or reveals certain properties based on the target model's output (Rigaki & Garcia, 2023). Our inference techniques work with any dataset or model type. We tested them on neural networks and black-box models from Amazon ML and Google Prediction API. We didn't know the learning algorithms or model architectures used by Amazon and Google. For testing, we used realistic classification tasks with datasets of images, retail purchases, location traces, and hospital inpatient stays (Shokri et al., 2017a). This paper shows inference attacks on three types of models: two from cloud-based services and one local model. They kept all the models’ black boxes, so they did not know their type, structure, or training specifics. The experiment shows that machine-learning-as-a-service platforms, like Google's and Amazon's, have the potential to leak significant information about their training data. For multi-class classification models trained on 10,000-record retail transaction datasets, membership inference attacks achieved median accuracy of 94% for Google and 74% for Amazon. Even without prior assumptions about the training data distribution and using fully synthetic data for shadow models, the accuracy against Google-trained models was still 90%. These findings highlight the need to address privacy vulnerabilities in machine learning services. This study discovered that machine learning services from Google and Amazon create models without warning users about privacy risks. Overfitting, the number of classes, and the amount of training data can cause information leakage. These services should consider these risks, inform customers, and provide ways to reduce data leakage. This study also quantifies how much information about the training data can be leaked through the model’s prediction without knowing the inner architecture of the model. This is achieved through black-box access to the target model, shadow training to mimic the behaviors of the target model and attack model, and an attack model (a classifier) to distinguish between the outputs of the shadow models on their training data ("members") versus their outputs on data they were not trained on ("non-members").

**Mitigation Strategy**: There are several ways a FL model can be attacked by a membership inference attack. We have found some key keyways to mitigate this type of attack. One effective approach is limiting the predictions vector to the top k classes, which is done by modifying the model's last layer to only output the probability of the k most likely classes, rather than all possible classes. The last layer of a machine learning model is chosen for this modification because it is the layer that produces the final output, which is the prediction vector. This layer contains the probabilities for each class based on the input data. By applying the filter at this stage, we can directly control and limit the information of the model outputs. This means that the process of identifying the most likely classes incorporates noise in a way that protects the privacy of the data used to train the model while still providing a useful indication of the most probable classes.

Another defense technique is coarsening the precision of the prediction vector: a strategy to minimize the risk of information leakage from machine learning models. This is done by making the probabilities in the prediction vector less precise, which means rounding or simplifying the numbers. For example, if a model is trying to identify whether an image is of a cat, dog, or rabbit, the prediction vector might look like this: [0.7, 0.2, 0.1]. This means the model predicts a 70% chance the image is a cat, a 20% chance it's a dog, and a 10% chance it's a rabbit.

A table with numbers and text

AI-generated content may be incorrect.

TABLE 1: This table compares the impact of different privacy-preserving techniques on a machine learning model's general prediction accuracy and its vulnerability to attacks on a hospital dataset (Shokri et al., 2017a).

Instead of providing highly detailed probabilities (like 0.8325 for a class), the model might round these numbers to fewer decimal places (like 0.83 or even 0.8). By doing this, the model reveals less specific information, making it harder for an attacker to infer details about the training data. For example, a prediction vector contains probabilities for different classes, while a feature vector contains numerical values representing the characteristics of an input.

One more method is to increase the entropy of the prediction vector, which makes the target model probabilities more spread out and uncertain. In the context of machine learning and information theory, entropy measures the unpredictability of the model's predictions. Higher entropy means the predictions are more spread out and uncertain, while lower entropy indicates more confident and concentrated predictions. For example, if a model predicts the probabilities for different classes as [0.32, 0.32, 0.31], the entropy is high because the probabilities are quite uniform and uncertain. Conversely, if the probabilities are [0.9, 0.05, 0.05], the entropy is low because the model is very confident about one class. This process can be done by modifying the softmax layer, which converts the model's raw scores (logits) into probabilities: increase its normalizing temperature t>0. This layer converts raw scores (logits) into probabilities. By increasing the temperature parameter t in the softmax function, we make the probabilities more spread out and uncertain. Applies the formula softmax to convert these scores into probabilities.

To prevent overfitting in machine learning, we use regularization techniques. One common method is L2 regularization, which discourages large parameter values by adding a penalty term to the model's loss function. This term is ​, where  are the model's parameters, and λ is the regularization factor. The larger the value of *λ*, the stronger the regularization effect during training, helping the model generalize better by keeping the parameters smaller (Shokri et al., 2017a).

**Evaluation of mitigation strategies:**

Liu, L., Wang, Y., Liu, G., Peng, K., & Wang, C. (2023). Membership Inference Attacks Against Machine Learning Models via Prediction Sensitivity. *IEEE Transactions on Dependable and Secure Computing*, *20*(3), 2341–2347. https://doi.org/10.1109/TDSC.2022.3180828

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017a). Membership Inference Attacks Against Machine Learning Models. *Proceedings - IEEE Symposium on Security and Privacy*, 3–18. https://doi.org/10.1109/SP.2017.41

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017b). Membership Inference Attacks Against Machine Learning Models. *Proceedings - IEEE Symposium on Security and Privacy*, 3–18. https://doi.org/10.1109/SP.2017.41