ACCURACY-PRIVACY TRADE-OFF IN DEEP ENSEMBLE: A MEMBERSHIP INFERENCE PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep ensemble learning has been shown to improve accuracy by training multiple neural networks and fusing their outputs. Ensemble learning has also been used to defend against membership inference attacks that undermine privacy. In this paper, we empirically demonstrate a trade-off between these two goals, namely accuracy and privacy (in terms of membership inference attacks), in deep ensembles. Using a wide range of datasets and model architectures, we show that the effectiveness of membership inference attacks also increases when ensembling improves accuracy. To better understand this trade-off, we study the impact of various factors such as prediction confidence and agreement between models that constitute the ensemble. Finally, we evaluate defenses against membership inference attacks based on regularization and differential privacy. We show that while these defenses can mitigate the effectiveness of the membership inference attack, they simultaneously degrade ensemble accuracy. The source code is available in supplementary materials.

1 INTRODUCTION

Ensemble learning has been shown to improve classification accuracy of neural networks in particular, and machine learning classifiers in general Kondratyuk et al. (2020); Kuncheva & Whitaker (2003); Sagi & Rokach (2018). The most commonly used approach in deep ensemble learning involves averaging the output of multiple neural networks (NN) that are independently trained on the same dataset with different random initialization Lobacheva et al. (2020). Such a simple approach has been extensively used in practice to improve accuracy Lee et al. (2015); Wang et al. (2020). Notably, a majority of the top performers in machine learning benchmarks, such as the ImageNet Large Scale Visual Recognition Challenge Russakovsky et al. (2015), have adopted some form of ensemble learning Lee et al. (2015); Szegedy et al. (2015); He et al. (2016).

Ensemble learning has also been used to defend against privacy-harming membership inference attacks, where the goal of the attacker is to infer whether a sample has been used to train a model–i.e., the sample belongs to the train set. Membership inference attacks generally use the prediction confidence of NN models to infer membership status of a sample Salem et al. (2018); Shokri et al. (2017); Truex et al. (2019); Yeom et al. (2018) by leveraging the insight that trained models may output higher prediction confidence on train samples than non-train samples Choo et al. (2020). The intuition behind using ensemble learning to defend against membership inference attacks is that training each model on a different subset of data makes the ensemble less prone to overfitting Salem et al. (2018). This idea of using ensemble learning to defend against membership inference attacks has since been discussed in the literature Huang et al. (2020); Li et al. (2021); Rahimian et al. (2020); Yang et al. (2020).

In this paper, we demonstrate the trade-off between accuracy and privacy in deep ensemble learning. Figure 1 illustrates this trade-off by plotting accuracy and membership inference attack effectiveness for ensembles comprising of varying number of base models (1, 2, 5, and 10) that are trained for different number of epochs (5, 45, and 90). We make two key observations here. First, there is an increase in both accuracy and membership attack effectiveness as we go from a single model to ensembles comprising of increasing number of base models. The trade-off is more noticeable for more accurate models that are trained for larger number of epochs. Second, we can adapt the design of ensembles to suitably navigate the trade-off between accuracy and privacy. Specifically, the two extreme cases are to: (1) maximize accuracy by using an ensemble of models at the same

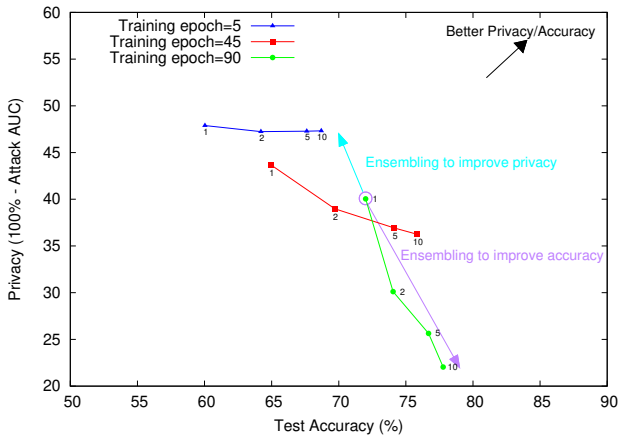


Figure 1: Trade-off between accuracy and privacy on an AlexNet model trained on CIFAR10. Each curve contains four points corresponding to ensembles comprising of 1, 2, 5, and 10 base models (from left to right). Using the single model trained for 90 epochs as a baseline, there are two choices: (1) making an ensemble of these models to achieve the highest accuracy possible but worse privacy (purple arrow), or (2) making an ensemble of less overfitted models (epoch #5) to achieve slightly lower accuracy of a single model but better privacy (cyan arrow).

level overfitting but at the cost of worse privacy (purple arrow); and (2) maximize privacy by using an ensemble of under-fitted models instead of a single model but at the cost of slight worse accuracy (cyan arrow).

To understand the root cause of this trade-off, we show that using ensemble learning to improve accuracy exacerbates its susceptibility to membership inference attacks by making train and non-train samples more distinguishable. By analyzing the confidence averaging mechanism of deep ensembles, we investigate all potential factors that enable membership inference. We show that the most influential factor is the level of correct agreement among models. Simply put, the number of models that correctly classify a train sample is often larger than the ones that correctly classify a test sample. This results in a wider confidence gap between train and non-train samples, when confidence values are averaged, enables more effective membership inference attacks.

We further show that the difference in the level of correct agreement between train and non-train samples stems from models' generalization gap. Hence, a natural question to ask is "can deep ensemble that use less overfitted models to mitigate privacy issues achieve high accuracy?". To answer this question, we study several regularization techniques, common membership inference defenses, and a few other ensembling approaches. We again observe a privacy-accuracy trade-off pattern similar to that shown in Figure 1.

Summary of contributions: In this paper, we perform a rigorous empirical study of membership inference attacks on deep ensemble models. First, we show that whenever the ensemble improves accuracy, it also leads to differences in the prediction confidence, which in turn enables more effective membership inference. Second, we analyze various factors that potentially contribute to this confidence change. We analyze various factors that potentially cause the prediction confidence of train and non-train samples to diverge and, consequently, allow more effective membership inference attacks: prediction confidence of correctly classifying models, prediction confidence of misclassifying models, and level of correct agreement. Among these factors, we show that the most dominant factor is the level of correct agreement among models which is correlated to the generalization gap of each model in an ensemble. We show that common defense mechanisms in membership inference literature, including differential privacy, L1 and L2 regularization, as well as other ensemble training approaches, such as bagging and partitioning, can be used to mitigate effectiveness of membership inference attacks but at the cost of accuracy.

In summary, we provide empirical evidences and analyses to demonstrate the trade-off between accuracy and privacy in deep ensemble models.

2 BACKGROUND

2.1 ENSEMBLE LEARNING

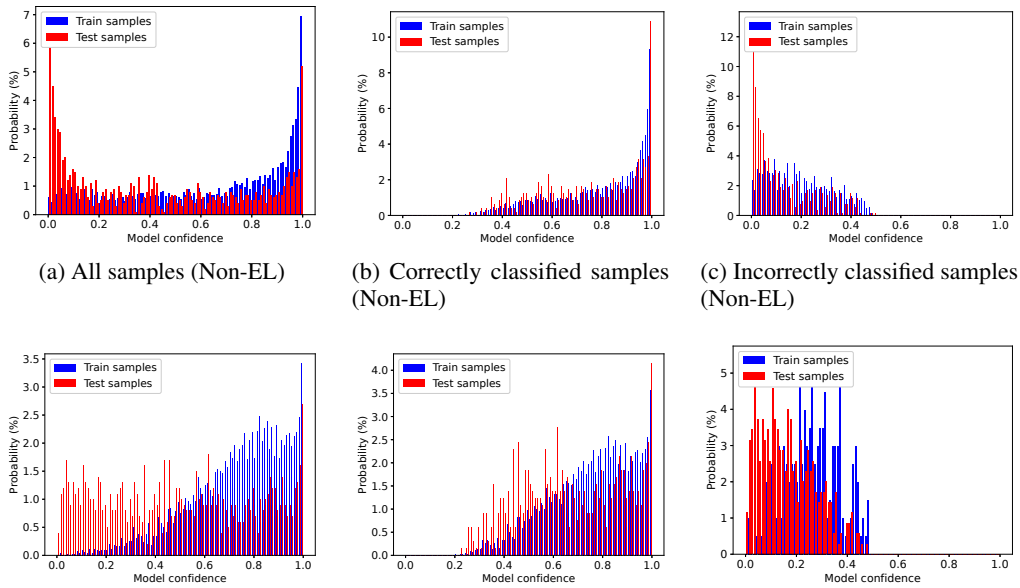
In literature, ensemble learning refers to various approaches that combine multiple models to make a prediction. Models used to construct an ensemble are often called base learners. There are two main factors to construct an ensemble Sagi & Rokach (2018): 1) how base learners are trained to ensure diversity, such as random initialization, bagging, partitioning, etc., and 2) how the output of base learners are fused to obtain the final output, including majority voting, confidence averaging, stacking, etc. Unlike ensemble of traditional machine learning algorithms, in a deep ensemble, the main source of diversity often comes only from random initialization of base learners Fort et al. (2019). In fact, other sources of diversity, such as bagging, have been shown to considerably degrade the overall accuracy of a deep ensemble Lee et al. (2015); Lakshminarayanan et al. (2017). We study bagging and partitioning as defense mechanisms because they prevent the exposure of all base learners to all training samples. In deep ensemble literature, the most common approach to obtain the output is to average the prediction confidences of NNs[1]. In this paper, we use ensemble learning (EL) and deep ensemble to refer to ensemble of deep neural networks trained with random initialization and fused with averaging the prediction confidence unless otherwise specified.

2.2 MEMBERSHIP INFERENCE

Membership inference is a form of privacy leakage where the goal is to determine if a sample was used as a part of a train set of a target model. The samples used during training are often referred to as member or train samples, and other samples are referred to as non-member, non-train, or test samples. The first membership inference attack on neural networks was proposed in Shokri et al. (2017) where the attacker trains an attack classifiers to predict the membership status. The attack classifiers take the prediction confidence of a target model as an input. Assuming that the attacker has access to a dataset with similar distribution, he/she trains a set of shadow models to mimic the target model. Since the membership status of the data with which the shadow models are trained are known to the attacker, he/she can use the data to train the attack classifier. Many papers use the same idea with different variations or less restrictive assumptions Salem et al. (2018); Liu et al. (2019); Song et al. (2019); Long et al. (2017); Truex et al. (2019); Long et al. (2018); Yeom et al. (2018); Rezaei & Liu (2021); Zou et al. (2020); Li & Zhang (2020). Most previous work built upon the idea of using prediction confidence to infer the membership status, except for Rezaei & Liu (2021); Choo et al. (2020); Rahimian et al. (2020). In Rezaei & Liu (2021), the authors assumed white-box access to the target model and launched a series of membership inference attack based on confidence values as well as distance to the decision boundary, gradient w.r.t model weight, and gradient w.r.t input. In Choo et al. (2020), the authors proposed two attacks based on input transformation and distance to the decision boundary in a black-box setting. Similarly, in Rahimian et al. (2020), the attacker randomly perturbs an input to obtain a set of random transformation of the input and uses the predicted label to infer membership status.

In this paper, we mainly focus on the most common form of membership inference attack based on confidence values. Recently, several studies report a simple baseline attack called *gap attack* Choo et al. (2020), also known as *naive attack* Rezaei & Liu (2021); Leino & Fredrikson (2020) that achieves similar performance in most scenarios. The gap attack predicts a sample as member if it is correctly classified by the target model, and as non-member otherwise. In other words, gap attack essentially reflects the generalization gap Rezaei & Liu (2021). In Rezaei & Liu (2021), authors extensively analyzed this phenomena in deep models and argued current membership inference attacks that barely outperform gap attacks are ineffective in practice because they only reflect the generalization gap and cannot infer the membership status of each individual sample accurately. Nevertheless, in this paper, we show that membership inference attacks on ensembles, unlike non-ensemble learning setting, significantly outperform the gap attack and, thus, pose serious threat to the privacy of training data.

[1]The less common approach is to average model logits which has been used in a few studies Webb et al. (2020); Wang et al. (2020). See Appendix A.3 for experimental evaluation of logit averaging. Throughout our experiments, we observed that both these approaches yield similar trends. However, the logit averaging seems to leak less membership status in some cases.



(a) All samples (Non-EL)

(b) Correctly classified samples (Non-EL)

(c) Incorrectly classified samples (Non-EL)

(d) All samples (EL-10)

(e) Correctly classified samples (EL-10)

(f) Incorrectly classified samples (EL-10)

Figure 2: AlexNet on CIFAR10 (Class #2): Confidence distribution of a single model (top row) versus an ensemble of 10 models (bottom row).

2.3 Membership Inference Defenses

Defense mechanisms against membership inference attacks can be summarized into two categories Rahimian et al. (2020):

Generalization-based: Shokri Shokri et al. (2017) was first to correlate membership inference success with overfitting. Since then, many standard regularization techniques have been used to alleviate overfitting, such as L1 regularization Choo et al. (2020), L2 regularization Choo et al. (2020); Truex et al. (2019); Nasr et al. (2019); Jia et al. (2019); Shokri et al. (2017), differential privacy Choo et al. (2020); Rahimian et al. (2020), dropout Jia et al. (2019), and adversarial training Nasr et al. (2018). Interestingly, ensemble learning has also been proposed as a defense mechanism. In Salem et al. (2018), they proposed a combination of partitioning and stacking as a defense mechanism. The intuition is that training each model with different subset of data makes the entire ensemble model less prone to overfitting. In this paper, we investigate bagging and partitioning as defense mechanisms. Note that these defense mechanisms often degrade the accuracy of the target model (see Section 4.2) Choo et al. (2020).

Confidence-masking: These defense mechanisms aim to reduce the amount of information that can be obtained from the output of a target model by perturbing Jia et al. (2019) or limiting the dimensionality of the output Shokri et al. (2017); Truex et al. (2019); Choo et al. (2020). Majority of these approaches are post-training methods and only work in black-box setting. Moreover, if adopted in ensemble learning settings, these approaches directly manipulate the output of ensemble averaging operation which is the core functionality of deep ensemble. In this paper, we mainly focused on generalization-based defenses.

3 How Does Ensembling Increase Membership Inference Effectiveness?

3.1 Confidence Distribution Shift

Ensemble learning is only helpful when base learners disagree on some samples Kuncheva & Whitaker (2003); Sagi & Rokach (2018). Otherwise, ensembling does not improve accuracy. Furthermore, when deep ensemble is used, the confidence values of multiple models are averaged to obtain the final prediction. Consequently, when ensembling improves accuracy, it averages the prediction confidence of highly confident predictions (mostly from models which correctly classified the sample) and less confident predictions (mostly from models which misclassified the sample). As a result, confidence distribution shift is inevitable for both train and test samples. This phenomenon can be observed as the distribution of Figure 2(a) shifts to that of 2(d). This can be better observed by separating correctly classified samples which have significantly higher prediction confidence (Figure 2(e)) and misclassified samples which have lower prediction confidence (Figure 2(f)). One can clearly observe that both distributions shift more towards the center from Figure 2(b) and (c) to Figure 2(e) and (f). However, the change in the distribution of train and test samples does not necessarily cause a more effective membership inference if the change has a similar effect on the confidence distribution of both train and test samples. In the next subsection, we analyze the potential factors that affect the distribution change and how they can change confidence distribution of member and non-member sets differently.

3.2 Effect of ensembling on Individual Samples

We use y_i to denote the confidence value of the ith model in an ensemble of n models. Hence, for a given sample x, the output of the ensemble is:

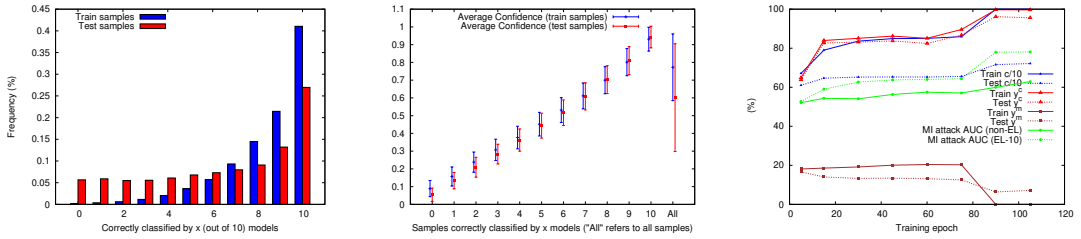
$$y_{el}(x) = \frac{\sum^n y_i(x)}{n} = \frac{\sum^c y_i^c(x) + \sum^m y_i^m(x)}{n}. \tag{1}$$

Given a single sample, we can further divide models in the ensemble into two groups: 1) models that correctly classified the sample denoted by y_i^c and 2) models that misclassified the sample by y_i^m. For a given sample x, c models correctly classify it and m models misclassify, where $c + m = n$. Note that the value of c and m depends on the sample [2].

Based on the Eq. (1), there are three major factors affecting the final confidence value (y_{el}) of a sample: y^c, y^m, and c. As a result, if these values are different for train and test samples, the ensembling causes different shift in the distributions, and consequently, membership inference attack will be more effective. These factors are as follows:

1. Confidence value of correctly classifying models (y^c): Since the majority of samples are supposed to be correctly classified by a practical model, any distinguishable confidence difference between train and test samples can lead to a very effective membership inference attack. However, as shown in Figure 2(b), we can observe that there is no significant difference between train and test samples.
2. Confidence value of misclassifying models (y^m): Unlike correctly classified samples, there is a marginal difference between confidence distribution of train and test samples of misclassified samples (see Figure 2(c)). This may be exploited by membership inference attack to partially distinguish between train and test samples.
3. Level of correct agreement (c) among models: As shown in Figure 3(a), the number of models that correctly classify a sample (c) is greater for train samples than test samples (see Appendix A.2 for more results). This is a direct reflection of generalization gap that causes more train samples to be correctly classified. Since prediction confidence of correctly classified samples are higher than misclassified samples on average, i.e., $y^c > y^m$, and c is smaller for test samples, the ensemble confidence of test samples (y_{el}) becomes lower than train samples. As a result, this factor can largely contribute to the effectiveness of membership inference attacks on deep ensembles.

[2] By an abuse of notation, we use c (m) to refer to (in)correctly classifying models and also as a superscript for the model output of (in)correctly classifying models, that is, y_i^c (y_i^m).



(a) Correct agreement distribution

(b) Distribution difference at each agreement level

(c) The effect of three factors on MI attack

Figure 3: AlexNet model trained on CIFAR10. Left: The distribution of the number of times a sample is correctly classified by 10 models used in the ensemble. The models often make less classification mistake on train samples that test samples. Center: By separating samples based on how many times they have been correctly classified, we can observe that the confidence output of these samples are negligible between train and test sets. Only when all samples are compared the distribution difference is significant and that is the direct effect of the third factor, namely the level of correct agreement. Right: The effect of the three factors on the MI attack. As the gap between the level of correct agreement of train and test widens (the blue lines), the MI attack on ensembles becomes more effective than a single model (green lines).

We note that the first two factors are not unique to ensembles and can be exploited by an attacker in a single model scenario as well. As a result, these two factors have been extensively studied in Rezaei & Liu (2021) in a single model scenario across various image datasets and well-trained models. They have shown that for deep models the first factor (y^c) is almost indistinguishable between train and test set and only the second factor (y^m) is marginally distinguishable. However, this marginal difference does not have a considerable impact on the different distribution shift in train and test sets. On the other hand, the level of agreement has a big impact on different distribution shifts. To better demonstrate this, we can analyze the distribution difference in each level of agreement separately. As shown in Figure 3(b), the average confidence of train and test samples are very close and indistinguishable when each level of agreement is drawn separately. If the effect of the first two factors were considerable, the two confidence values for each level of agreement would have been more distinguishable. Note that the average confidence between train and test sets is more distinguishable for the first two points in x-axis (where majority of models misclassify a sample). However, these distributions only constitute a tiny portion of the training dataset, as shown with the first two blue bins in Figure 3(a). However, when all samples are combined, we can vividly observe that the average confidence gap between train and test sets considerably widens, as shown in the last point in x-axis in Figure 3(b). This clearly demonstrates that the major factor in different distribution shift between train and test sets is the level of agreement (c).

Note that, unlike the first two factors, the third factor (c) only improves the effectiveness of membership inference attacks in ensemble scenario because it does not exist in a single model. In other words, if a defence strategy eliminates the effect of the average level of correct agreement (i.e., it ensures that c is close between train and test samples), the membership inference attack is still possible on the ensemble, but only to the degree that it is possible on a single model[3]. As shown in Figure 3(c), as the gap between y^c of train and test sets (red lines) and the gap between y^m of train and test sets (brown lines) increases, the attack on both single model (non-ensemble) and also the ensemble (EL-10) increases. However, only when the average level of correct agreement gap between train and test (blue lines) widens, the membership inference attack on ensembles becomes more effective than on non-ensembles.

Another important observation from Figure 3(c) is that the minimum level of agreement gap between train and test occurs when models are relatively underfitted (i.e., the blue lines in first few epochs).

[3]Although this can be understood by analyzing the Eq. (1), it is difficult to demonstrate empirically. The reason is that these three factors are not independent, and, hence, our attempts to significantly change the third factor without changing the other two factors have been unsuccessful. See Appendix A.2 to see the dependency between these two factors across all datasets and models.

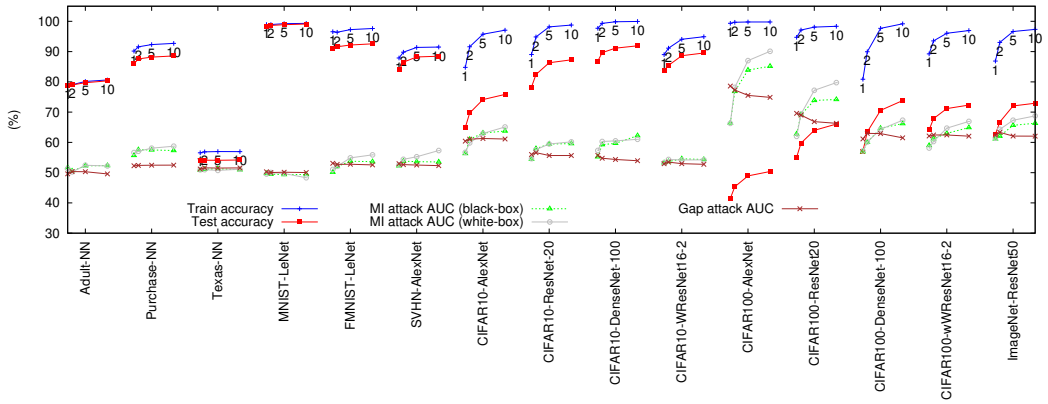


Figure 4: Membership inference attack results across all datasets/models. Each curve indicates an ensemble of 1 (non-EL), 2, 5, and 10 models from left to right.

This phenomena has also been partially observed in Fort et al. (2019) (Figure 2(c)). The main reason is that underfitted models often only learn the most common and generalizable features and, thus, they often agree on the features and predictions. As they move from underfitted region to overfitted region, they start to memorize features more specific to train samples Feldman (2020). Consequently, the average gap of correct agreement between train and test set widens. Hence, the wider generalization gap of base learners is, the more effective membership inference attack would be on ensemble.

3.3 Why Does it Outperform Gap Attack Significantly?

Figure 4 shows that the effectiveness of membership inference attacks increases and outperforms the gap attack as ensemble is deployed. This raises significant privacy concern since the gap attack is often suggested as a baseline that is also hard to outperform in non-ensemble setting Rezaei & Liu (2021); Leino & Fredrikson (2020); Choo et al. (2020). Note that gap attack can be viewed as a metric directly reflecting the generalization gap rather than a reliable membership inference. As suggested in Rezaei & Liu (2021), we can separate correctly classified samples and misclassified samples to understand why membership inference attacks can barely outperform gap attack. As shown in Figure 2(b) and (c), the distributions of train and test samples are similar when separated into correctly classified and misclassified sets. As elaborated in Rezaei & Liu (2021), the reason why the distribution of all samples (Figure 2(a)) looks more distinguishable when correctly classified samples and misclassified samples are combined is that there are often more misclassified samples in the test set than the train set. This is the information that gap attack exploits which essentially reflects the generalization gap. In order for a membership inference attack to considerably outperform the gap attack, the distribution of correctly classified samples and misclassified samples should leak membership status information, which is not often the case as it is shown in Figure 2(b) and (c), and Rezaei & Liu (2021).

When ensembling is used, the distribution of confidence values changes dramatically, as explained in Section 3.1. By comparing the confidence distribution of correctly classified samples in an ensemble (Figure 2(e)) with a non-ensemble scenario (Figure 2(b)), the distribution is clearly more distinguishable in ensemble case. This is of significant privacy concern because, as discussed in Rezaei & Liu (2021), majority of samples in practice belong to the correctly classified set. Similar trend is also observable in misclassified samples (Figure 2(f)). Hence, the confidence values, that barely leak more information than generalization gap itself in a single model scenario, now considerably leak more membership information than just the generalization gap. That is the reason why membership inference attacks are significantly more effective in ensembles in comparison to the gap attack.

4 EXPERIMENTS RESULTS

4.1 EXPERIMENTAL SETUP

We explore a wide range of datasets that are often used in deep ensemble literature or membership inference literature: Adult[4], Texas, Purchase Shokri et al. (2017), MNIST LeCun et al. (1998), FMNIST Xiao et al. (2017), SVHN Netzer et al. (2011), CIFAR10 Krizhevsky et al. (2009), CIFAR100 Krizhevsky et al. (2009), and ImageNet Russakovsky et al. (2015). For non-image datasets (Adult, Purchase, and Texas), we use a fully connected neural network consisting of a hidden layer of size 128 and a Softmax layer. All other training parameters for these datasets are set as suggested in Shokri et al. (2017). For image datasets, we use a wide range of convolutional neural networks depending on the input dimension and the difficulty of the task. We use the model implementations adopted in Nasr et al. (2019); Rezaei & Liu (2021)[5]. We train 10 models for each dataset with random initialization and construct an ensemble of 2, 5 and 10 models. We train each model for up to 200 epochs.

For membership inference attack models, we consider two different threat models: black-box and white-box. In the black-box setting, the attacker only access to the output of the ensemble, which is the average over softmax output of individual models. In the white-box setting, the attacker can see the output of each model in ensemble. In the former case, the attack model takes the output of ensemble as input, while in the latter case, it takes all confidence outputs of all models as input. Attack models are NNs with three hidden layers of size 128, 128, and 64, respectively. We consider worst case scenario where 80% of the training dataset is given to the attacker and the goal is to infer the membership of the remaining samples, similar to Rezaei & Liu (2021). All other training parameters are set as suggested in Rezaei & Liu (2021). All experiments are implemented in Python 3.5, PyTorch 1.6 and are conducted on a server with Intel Xeon Processor 24-core 2.70GHz, 8 GeForce RTX 2080 GPUs and 384GB memory. See Appendix A.1 for full report.

Figure 4 shows the results on all datasets. For some datasets, such as Adult, Texas, and MNIST, ensemble learning barely changes the accuracy or privacy. That is because the disagreement across models is insignificant in these datasets. For all other datasets, ensemble learning improves the accuracy as well as the effectiveness of membership inference attacks. As mentioned in Section 3.2, the most salient factor in membership inference effectiveness on deep ensembles is the accuracy gap between train and test set. Figure 4 clearly shows that whenever this generalization gap is large for non-ensemble case, the attack improvement is significant after ensembling. It is worth nothing that the ensembling can often reduce the generalization gap and the effectiveness of the gap attack (e.g., CIFAR10-DenseNet-100, CIFAR100-AlexNet, or CIFAR100-ResNet20). However, due to the reasons explained in Section 3.3, the membership inference effectiveness still increases.

4.2 DEFENSE MECHANISMS

As discussed in previous sections, the main factor that causes more privacy leakage in ensemble learning is the generalization gap of the base models. Therefore, any standard regularization technique can potentially work as a defense mechanism. In this paper, we study L1 and L2 regularization, DP-SGD[6], and MMD+Mixup Li et al. (2021). Furthermore, one can simply terminate training early to keep the model's weights in less overfitted region. Moreover, some ensembling techniques, such as bagging and partitioning, limits the access of models to all training samples, which can potentially reduce membership inference effectiveness on deep ensembles.

We train 10 AlexNet models with different initialization on CIFAR10 for each defense mechanism. Here, we consider black-box membership inference attack to measure privacy. Figure 5 shows the effect of defense mechanisms on ensemble learning. We can observe a consistent trade-off between ensemble accuracy and privacy that resembles Pareto optimal points. Note that privacy degradation rate is clearly not a constant. An ensemble of heavily regularized models or under-fitted models barely causes more privacy leakage (e.g., L2 regularization with $\lambda = 1$). On the other hand, an ensemble of overfitted models (e.g., non-regularized models trained for 90 epochs) results in large privacy leakage. It is worth mentioning that for a given accuracy, some approaches provide more

[4]http://archive.ics.uci.edu/ml/datasets/Adult

[5]https://github.com/bearpaw/pytorch-classification

[6]We use Opacus implementation in PyTorch from https://github.com/pytorch/opacus.

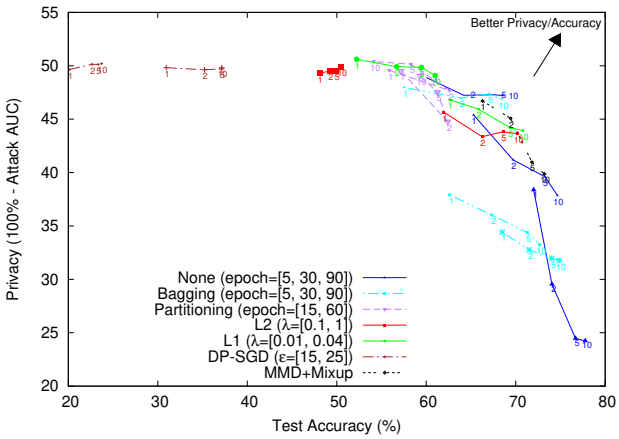


Figure 5: Effect of defense mechanisms on an AlexNet model trained on CIFAR10. There is a trade-off between accuracy and MI attack performance. The size of each point indicates the relative value of its parameter, i.e., epoch, λ and ϵ.

privacy. For instance, the regular deep ensemble and the bagging of models trained for 30 epochs achieve similar accuracy. However, bagging is in fact more prone to membership inference because each model in bagging is exposed to fewer training data samples and consequently is more overfitted.

5 Discussion

In this section, we list some limitations of our empirical analysis and opportunities for future work. First, privacy is a multi-faceted concept and can be defined or quantified in several ways. In this work, we quantified privacy leakage in terms of the effectiveness of membership inference attacks. Future work can quantify privacy in terms of other relevant attacks such as model inversion Fredrikson et al. (2015); He et al. (2019), property inference Ateniese et al. (2015); Ganju et al. (2018), and model stealing attack Tramèr et al. (2016) as well as formally-provable measures such as differential privacy Dwork (2008). Second, ensemble learning is an umbrella term covering a variety of methods to combine multiple base learners. An arbitrary method of training and combining base learners can be construed as ensemble learning. We limited the scope of our study to ensemble averaging because of its prevalent use for ensembling deep models. Given the privacy leakage of deep ensembles demonstrated here, future work can look into designing other fusion approaches to better navigate the trade-off between privacy and accuracy.

6 Conclusion

In this paper, we investigate membership inference attacks in deep ensemble learning. We empirically demonstrate that there is a trade-off between accuracy and privacy. Specifically, we show that the widely used deep ensemble technique, which involves random initialization and confidence averaging, increases accuracy but degrades privacy in terms of more effective membership inference. We study the effect of various regularization-based defense mechanisms as well as other ensembling techniques on the trade-off between accuracy and privacy. By analyzing the way ensemble averaging works, we identify various factors that contributes to the privacy leakage in deep ensembles. We show that the most salient factor is the level of agreement among base learners which can be mitigated by reducing the generalization gap. Although generalization gap does not always correlate with more effective membership inference in comparison to a baseline (gap attack) in a single model, it allows a significantly more effective membership inference attack in deep ensembles. As part of our work, we are interested in designing a more sophisticated ensembling approach that is more robust to membership inference attacks without sacrificing accuracy.

REFERENCES

Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

Christopher A Choquette Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. *arXiv preprint arXiv:2007.14321*, 2020.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pp. 1–19. Springer, 2008.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

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*, pp. 1322–1333, 2015.

Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 619–633, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 148–162, 2019.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Hongwei Huang, Weiqi Luo, Guoqiang Zeng, Jian Weng, Yue Zhang, and Anjia Yang. Damia: Leveraging domain adaptation as a defense against membership inference attacks. *arXiv preprint arXiv:2005.08016*, 2020.

Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 259–274, 2019.

Dan Kondratyuk, Mingxing Tan, Matthew Brown, and Boqing Gong. When ensembling smaller models is more efficient than single large models. *arXiv preprint arXiv:2005.00570*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6405–6416, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pp. 1605–1622, 2020.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pp. 5–16, 2021.

Zheng Li and Yang Zhang. Label-leaks: Membership inference attack with label. *arXiv preprint arXiv:2007.15528*, 2020.

Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, and Wenqing Cheng. Socinf: Membership inference attacks on social media health data with machine learning. *IEEE Transactions on Computational Social Systems*, 6(5):907–921, 2019.

Ekaterina Lobacheva, Nadezhda Chirkova, Maxim Kodryan, and Dmitry Vetrov. On power laws in deep ensembles. *arXiv preprint arXiv:2007.08483*, 2020.

Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*, 2017.

Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646, 2018.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. In *2019 ieee symposium on security and privacy*, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop*, 2011.

Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. Sampling attacks: Amplification of membership inference attacks by repeated queries. *arXiv preprint arXiv:2009.00395*, 2020.

Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 241–257, 2019.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pp. 601–618, 2016.

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019.

Xiaofang Wang, Dan Kondratyuk, Kris M Kitani, Yair Movshovitz-Attias, and Elad Eban. Multiple networks are more efficient than one: Fast and accurate models via ensembles and cascades. *arXiv preprint arXiv:2012.01988*, 2020.

Andrew Webb, Charles Reynolds, Wenlin Chen, Henry Reeve, Dan Iliescu, Mikel Lujan, and Gavin Brown. To ensemble or not ensemble: When does end-to-end training fail? *stat*, 1050:6, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Ziqi Yang, Bin Shao, Bohan Xuan, Ee-Chien Chang, and Fan Zhang. Defending model inversion and membership inference attacks via prediction purification. *arXiv preprint arXiv:2005.03915*, 2020.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018.

Shaofeng Zhang, Meng Liu, and Junchi Yan. The diversified ensemble neural network. *Advances in Neural Information Processing Systems*, 33, 2020.

Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning. *arXiv preprint arXiv:2009.04872*, 2020.

A Appendix

A.1 Full Training Report

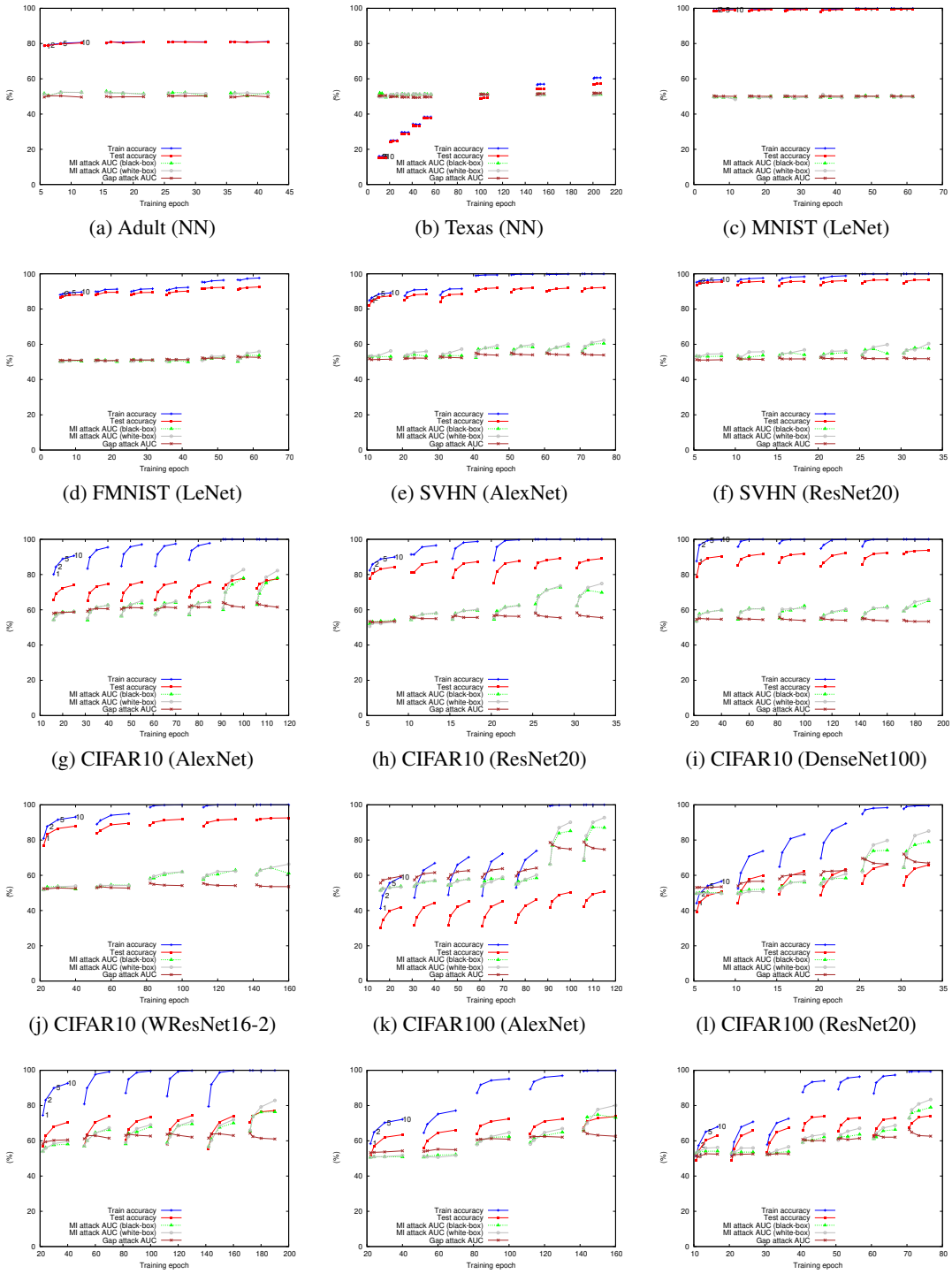
Figure 6 demonstrates the improvement of accuracy and MI attack over various training epochs. For datasets that ensembling outperforms a single model, using an ensemble of underfitted models is less prone to MI attack. However, it corresponds to lower accuracy. This is consistent with all experiments across different datasets and models, shown in Figure 6.

A.2 Level of Correct Agreement

As discussed in section 3.2, overfitted models tend to disagree more on test samples than train samples. In other word, the distribution of agreement for train and test sets becomes more distinguishable as models overfit. This distinction is more clear for datasets, such as CIFAR10 and CIFAR100, which shows most improvement when ensembling is used, as shown in Figure 7, 8, and 9. Furthermore, the level of agreement can reveal if an ensemble can actually improve prediction. If all models correctly classify a sample or all models misclassify a sample, ensembling fails to outperform a single model. This is the case for Adult, Texas, and MNIST, as shown in Figure 9.

A.3 Averaging Logit

In this section, we average logits (the output of a model before Softmax) of NN models instead of the confidences. We can still observe that ensembling leaks more membership status than non-ensemble scenario. However, the MI attacks with average confidence (Figure 6), in general, are slightly more effective than MI attacks with average logits (Figure 10). The reason is that that confidence values are normalized and, hence, when aggregated all models have the same contribution to the overall confidence output of the ensemble. However, when logit is used, the confidence output of the ensemble is more influenced by highly activated neurons. These highly activated neurons, which



(a) Adult (NN) (b) Texas (NN) (c) MNIST (LeNet)

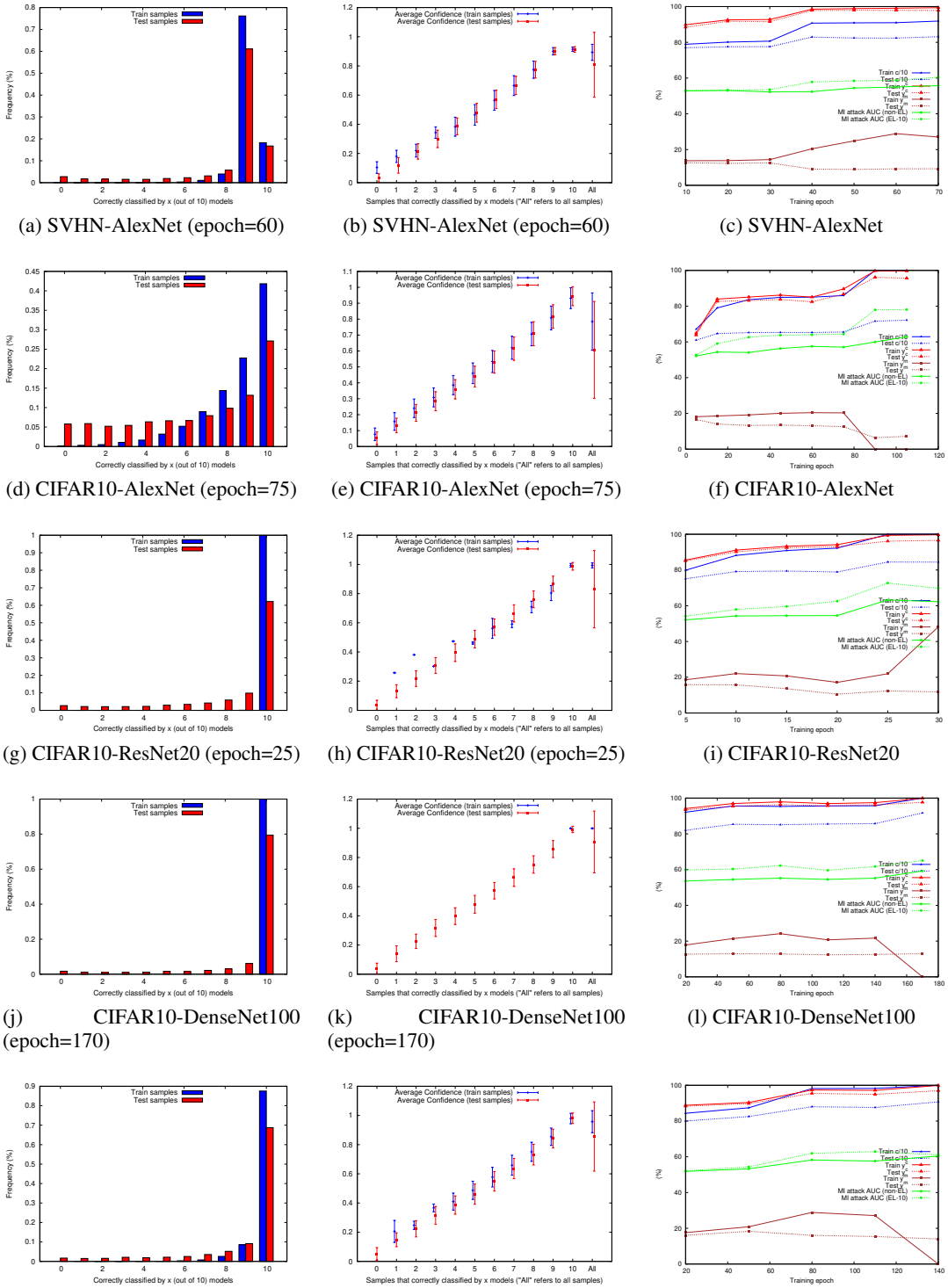
(d) FMNIST (LeNet) (e) SVHN (AlexNet) (f) SVHN (ResNet20)

(g) CIFAR10 (AlexNet) (h) CIFAR10 (ResNet20) (i) CIFAR10 (DenseNet100)

(j) CIFAR10 (WResNet16-2) (k) CIFAR100 (AlexNet) (l) CIFAR100 (ResNet20)

(m) CIFAR100 (DenseNet100) (n) CIFAR100 (WResNet16-2) (o) ImageNet (ResNet50)

Figure 6: Target models' accuracy and MI attacks' AUC across all datasets and models.



(a) SVHN-AlexNet (epoch=60) (b) SVHN-AlexNet (epoch=60) (c) SVHN-AlexNet

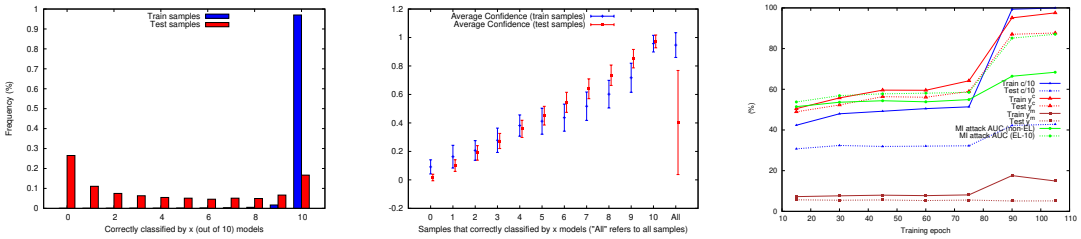
(d) CIFAR10-AlexNet (epoch=75) (e) CIFAR10-AlexNet (epoch=75) (f) CIFAR10-AlexNet

(g) CIFAR10-ResNet20 (epoch=25) (h) CIFAR10-ResNet20 (epoch=25) (i) CIFAR10-ResNet20

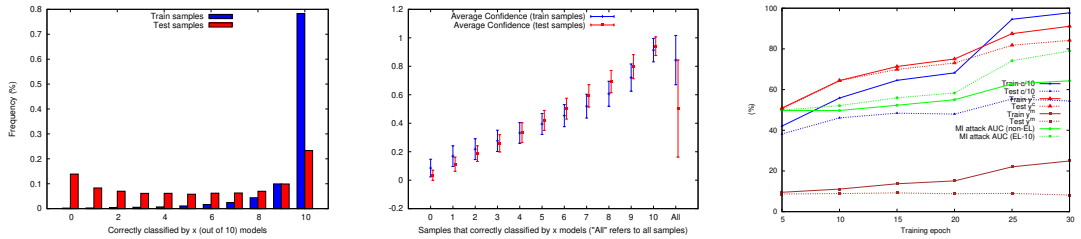
(j) CIFAR10-DenseNet100 (epoch=170) (k) CIFAR10-DenseNet100 (epoch=170) (l) CIFAR10-DenseNet100

(m) CIFAR10-WResNet16-2 (epoch=80) (n) CIFAR10-WResNet16-2 (epoch=80) (o) CIFAR10-WResNet16-2

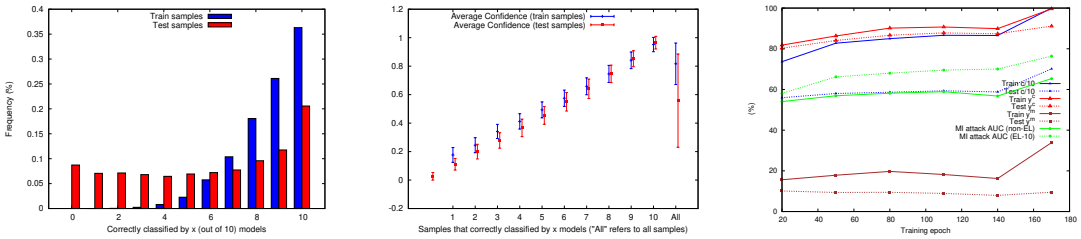
Figure 7: The first row contains correct agreement distribution for SVHN and CIFAR10 datasets. The second row shows the average and standard deviation of distribution of samples based on the level of correct agreement. The third row shows the effect of the three factors on the MI attack.



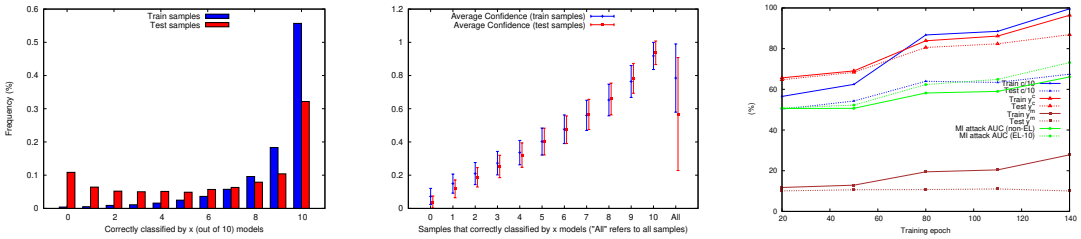
(a) CIFAR100-AlexNet (epoch=90) (b) CIFAR100-AlexNet (epoch=90) (c) CIFAR100-AlexNet



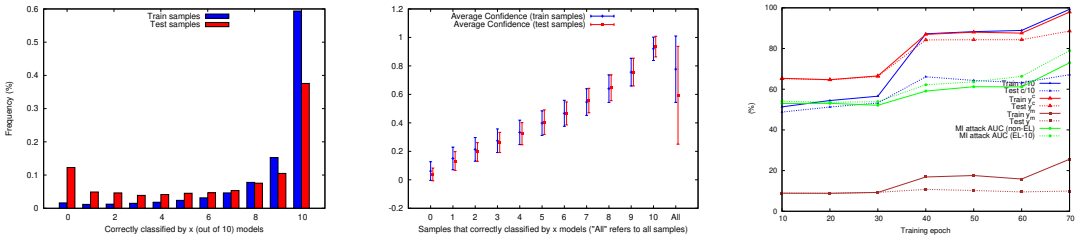
(d) CIFAR100-ResNet20 (epoch=25) (e) CIFAR100-ResNet20 (epoch=25) (f) CIFAR100-ResNet20 (epoch=25)



(g) CIFAR100-DenseNet100 (epoch=140) (h) CIFAR100-DenseNet100 (epoch=140) (i) CIFAR100-DenseNet100

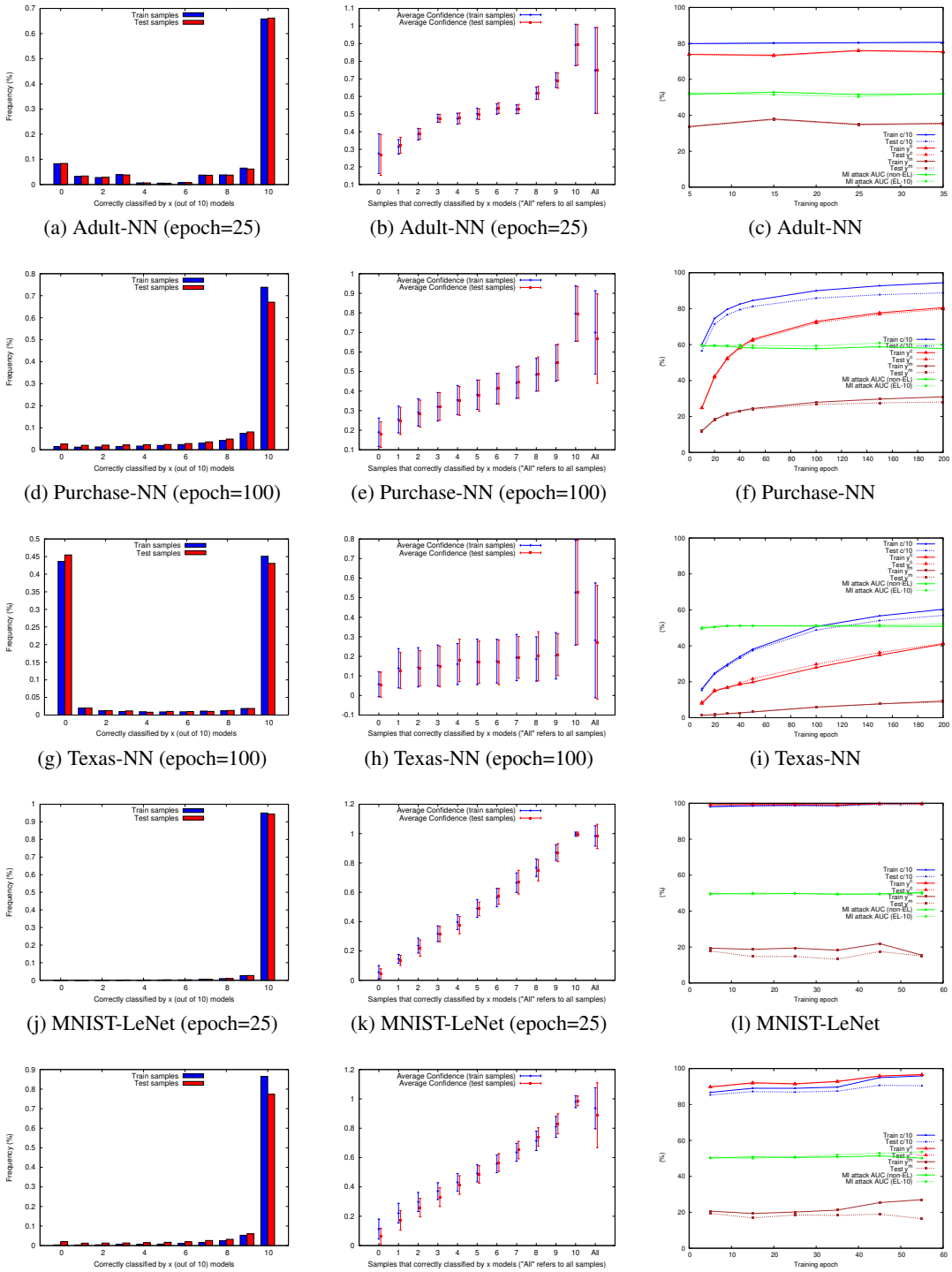


(j) CIFAR100-WResNet16-2 (epoch=110) (k) CIFAR100-WResNet16-2 (epoch=110) (l) CIFAR100-WResNet16-2



(m) ImageNet-ResNet50 (epoch=40) (n) ImageNet-ResNet50 (epoch=40) (o) ImageNet-ResNet50

Figure 8: The first row contains correct agreement distribution for CIFAR100 and ImageNet datasets. The second row shows the average and standard deviation of distribution of samples based on the level of correct agreement. The third row shows the effect of the three factors on the MI attack.



(a) Adult-NN (epoch=25) (b) Adult-NN (epoch=25) (c) Adult-NN

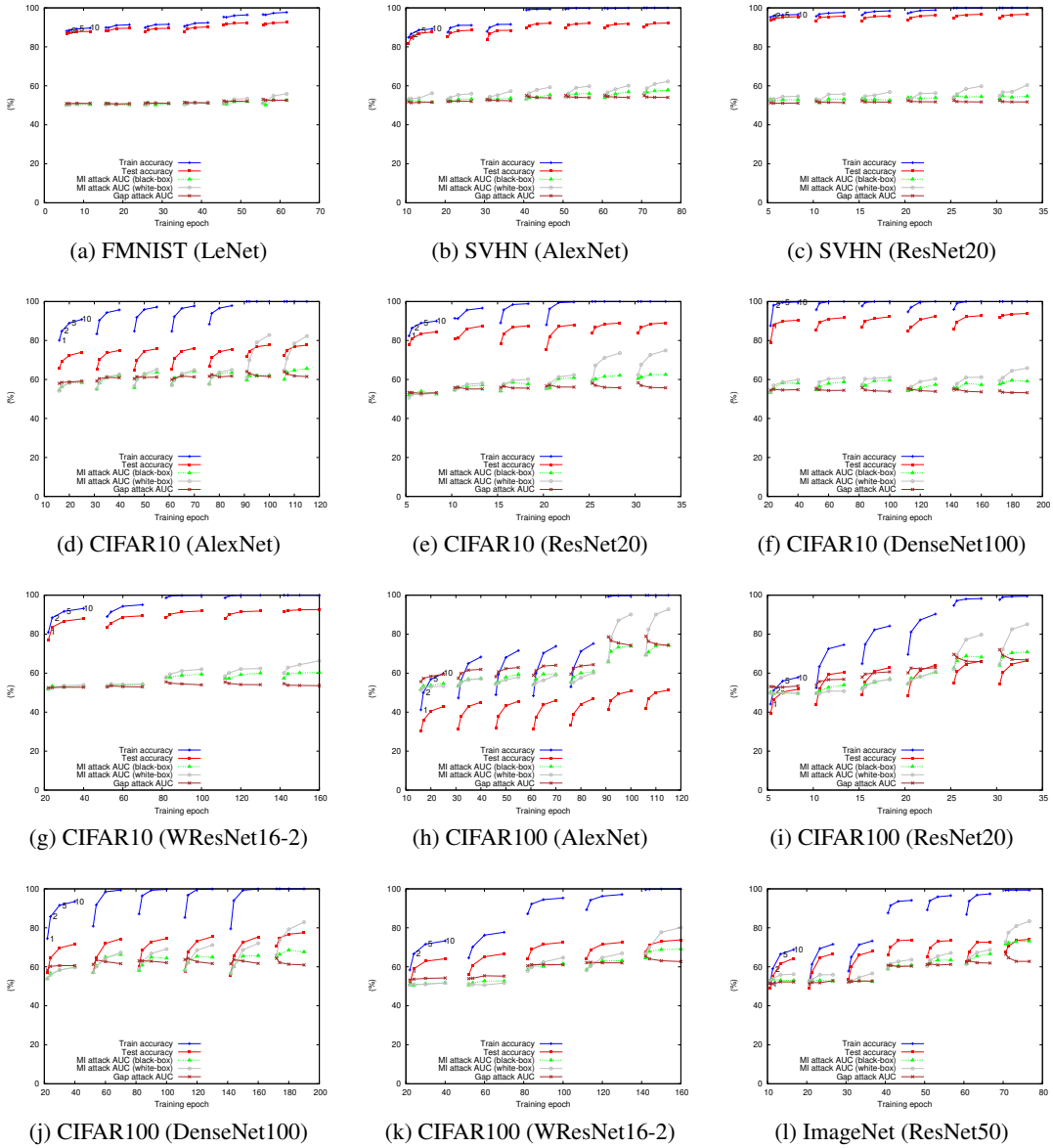
(d) Purchase-NN (epoch=100) (e) Purchase-NN (epoch=100) (f) Purchase-NN

(g) Texas-NN (epoch=100) (h) Texas-NN (epoch=100) (i) Texas-NN

(j) MNIST-LeNet (epoch=25) (k) MNIST-LeNet (epoch=25) (l) MNIST-LeNet

(m) MNIST-LeNet (epoch=55) (n) MNIST-LeNet (epoch=55) (o) MNIST-LeNet

Figure 9: The first row contains correct agreement distribution for Adult, Purchase, Texas, MNIST, and FMNIST datasets. The second row shows the average and standard deviation of distribution of samples based on the level of correct agreement.The third row shows the effect of the three factors on the MI attack.



(a) FMNIST (LeNet) (b) SVHN (AlexNet) (c) SVHN (ResNet20)

(d) CIFAR10 (AlexNet) (e) CIFAR10 (ResNet20) (f) CIFAR10 (DenseNet100)

(g) CIFAR10 (WResNet16-2) (h) CIFAR100 (AlexNet) (i) CIFAR100 (ResNet20)

(j) CIFAR100 (DenseNet100) (k) CIFAR100 (WResNet16-2) (l) ImageNet (ResNet50)

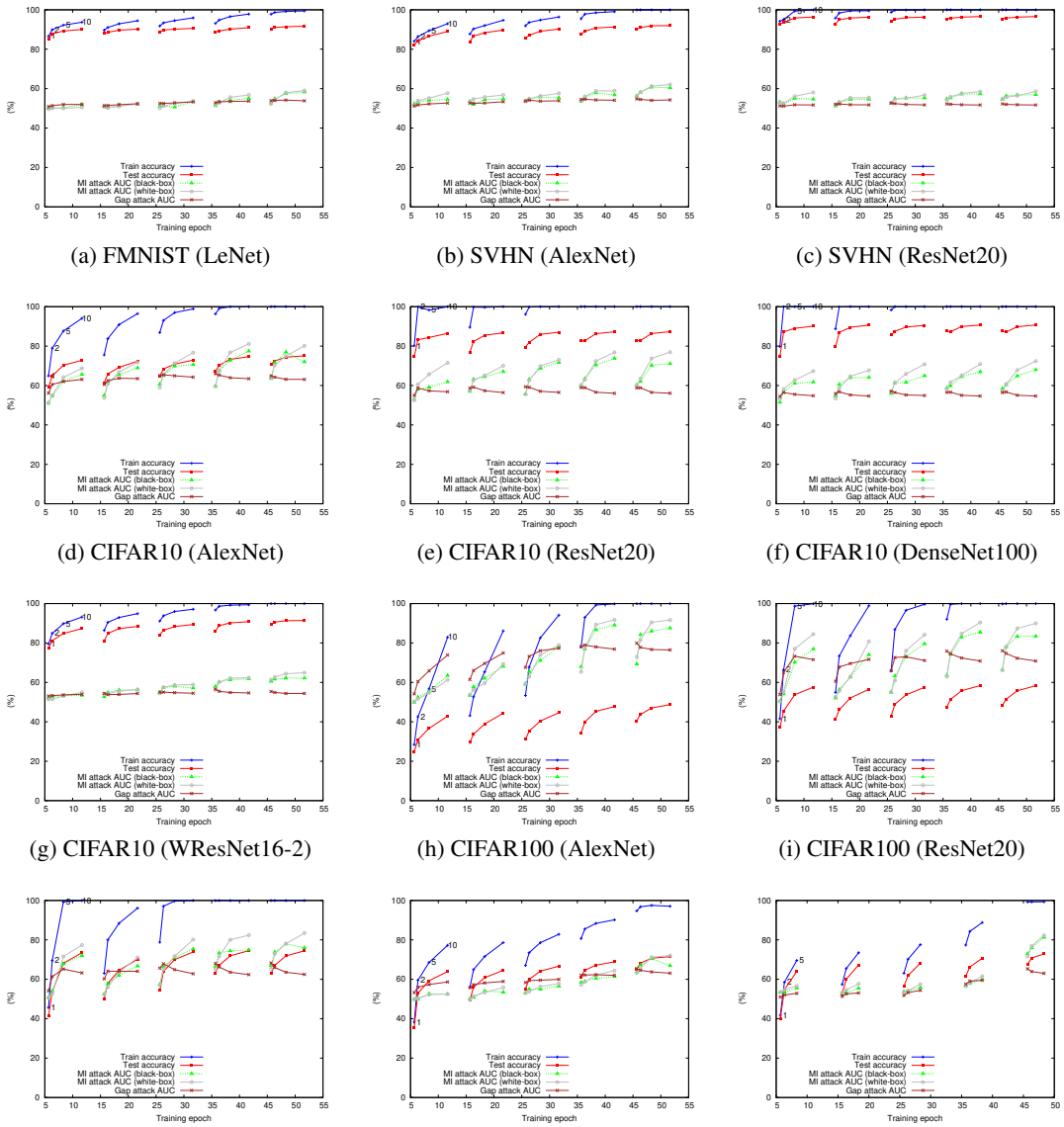
Figure 10: Target models' accuracy and MI attacks' AUC across all datasets and models. Here, MI attack model uses aggregated logits instead of aggregated confidences.

often belong to the correctly classifying models, has significantly more influence on the confidence output of ensemble in comparison with lightly activated neurons of misclassifying models. Hence, the confidence output is heavily influenced by only a portion of models in ensemble that have high activation neurons. In other words, it can be seen as an ensemble of only a portion of models, not all models in the ensemble. Since ensembling with fewer models leaks less membership status, logit averaging of n models leak membership status than confidence averaging of the same number of models. Note that logit averaging is still prone with the same degree to membership leakage in a white-box attack since a white-box attacker has access to all confidence values. Note that the consequence of using logit in certain applications, such as confidence estimation, that requires reliable confidence estimation is out of the scope of this paper. Moreover, one major drawback of using logit is that it can be arbitrary scaled Wang et al. (2020). However, in scenarios where only accuracy is concerned and white-box access is unavailable to the attacker, averaging logits seems to have a better privacy protection of training data than averaging confidences.

A.4 More Advanced Ensembling Approaches

In this section, we evaluate two state-of-the-art ensembling approaches, namely snapshot ensembles Huang et al. (2017) and diversified ensemble networks Zhang et al. (2020). For snapshot ensemble, we train several models on several datasets for 500 epochs and restart the cycle every 50 epochs, similar to the original paper Huang et al. (2017). Note that the goal of our evaluation is show the accuracy-privacy trade-off, not to achieve the highest accuracy possible. Due to this reason and limited time we had, we did not perform an exhaustive hyper-parameter tuning. Nevertheless, similar trade-off can be observed in Figure 11.

We also conduct the same experiment with diversified ensemble networks Zhang et al. (2020). The original paper used pre-trained VGG and ResNet models. We did not use pre-trained models for two reasons: 1) It make comparison with other approaches unfair, and 2) It may interfere with the membership inference analysis. We find that by using randomly initialized models to start training, L_d varies significantly and prevents the optimization to converge. Therefore, we add a weight to the L_d term to reduce its effect on the entire loss. We use 0.01 for CIFAR10 and SVHN and 0.001 for CIFAR100. For the shared layer, we use a fully-connected layer of size 128 followed by batch normalization and Relu activation. We use SGD to train models for 60 epochs while droping the learning at each 20 epoch by 0.1. We could not achieve the exact same results as reported in the paper for two main reasons: 1) we did not use pre-trained models in the ensemble, and 2) many hyper-parameters and implementation details are missed from the original paper.



(a) FMNIST (LeNet) (b) SVHN (AlexNet) (c) SVHN (ResNet20)

(d) CIFAR10 (AlexNet) (e) CIFAR10 (ResNet20) (f) CIFAR10 (DenseNet100)

(g) CIFAR10 (WResNet16-2) (h) CIFAR100 (AlexNet) (i) CIFAR100 (ResNet20)

(j) CIFAR100 (DenseNet100) (k) CIFAR100 (WResNet16-2) (l) ImageNet (ResNet50)

Figure 11: Target models' accuracy and MI attacks' AUC across all datasets and models using snapshot ensemble Huang et al. (2017).