Is EMA Robust? Examining the Robustness of Data Auditing and a Novel Non-calibration Extension

Ayush Alag
Princeton University
aalag@princeton.edu

Yangsibo Huang
Princeton University
yangsibo@princeton.edu

Kai Li
Princeton University
li@princeton.edu

Abstract

Auditing data usage in machine learning models is crucial for regulatory compliance, especially with sensitive data like medical records. In this study, we scrutinize potential vulnerabilities within an acknowledged baseline method, Ensembled Membership Auditing (EMA), which employs membership inference attacks to determine if a specific model was trained using a particular dataset. We discover a novel False Negative Error Pattern in EMA when applied to large datasets, under adversarial methods like dropout, model pruning, and MemGuard. Our analysis across three datasets shows that larger convolutional models pose a greater challenge for EMA, but a novel metric-set analysis improves performance by up to 5%. Orthogonally, we introduce EMA-Zero, a GAN-based dataset auditing method that does not require an external calibration dataset. Notably, EMA-Zero performs comparably to EMA with synthetic calibration data trained on as few as 100 samples.

1 Introduction

Federated learning (FL) (McMahan et al., 2017) is a collaborative learning framework that allows participants to collectively train a machine learning model without sharing their private data. A crucial aspect of FL is respecting participants' "right to be forgotten" (RTBF), a legal requirement mandated by regulations like the General Data Protection Regulation (GDPR) (Voigt and Von dem Bussche, 2017) and the California Consumer Privacy Act (CCPA) (Legislature, 2018), which enables participants to request data removal from trained models.

The concept of machine unlearning has arisen as a solution to eliminating data from machine learning models, and various machine unlearning methods have been proposed (Ginart et al., 2019; Liu et al., 2020; Wu et al., 2020; Bourtoule et al., 2021; Izzo et al., 2021; Sekhari et al., 2021; Gupta et al., 2021; Ye et al., 2022). However, from the perspective of an FL participant, a more immediate concern is *unlearning verification* or *data auditing*, the process of confirming whether their data has genuinely been removed from a model. Ensuring reliable data auditing forms the cornerstone of compliance with RTBF. Unfortunately, there are only a few works in the field of unlearning verification/data auditing (Huang et al., 2022; Weng et al., 2022; Gao et al., 2022; Zhou et al., 2023; Shi et al., 2023), and very limited work has been done to assess their robustness and practicality in real-world scenarios.

In this study, we focus on scrutinizing EMA (Huang et al., 2022), a plug-and-play unlearning verification method that does not require modifications during the training stage (as required by Weng et al. (2022); Gao et al. (2022); Zhou et al. (2023)), making it more practical. Our key contributions are:

- We systematically evaluate EMA's performance against adversarial methods and introduce improvements to EMA through a novel metric-set analysis (Section 3.2).
- We conduct ablation studies to understand EMA's poorer performance on larger-scale settings, with several hypotheses and error patterns explored (Section 3.3).
- We extend the capabilities of EMA by introducing EMA-Zero (Section 4), a variant designed to function in a no-calibration setting. EMA-Zero utilizes GANs to generate synthetic

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

data for shadow model training. This innovative approach brings EMA-Zero close to the baseline EMA's performance level while offering applicability in various collaborative learning scenarios.

2 Preliminary: Ensembled Membership Auditing (EMA)

Ensembled Membership Auditing (EMA) relies on the membership inference attack (Shokri et al., 2017), a process to determine whether an individual data point was included in the training of the target model. Specifically, EMA achieves data auditing by aggregating data-wise membership scores and applying statistical testing, using the following three steps:

1. Calibration Model Training: A calibration model with the same architecture as the target model is trained on a held-out dataset from the same distribution as the training data.
2. Per-Sample Metric Thresholding: Given calibration training and test datasets D_{ctr}, D_{ctest} and a metric m, the calibration model generates an independent threshold for m that best partitions D_{ctr} and D_{ctest}. The specific metrics utilized by baseline EMA are correctness, confidence, and negative entropy. These thresholds are then applied to the target model's outputs on each sample in the query dataset D_q, yielding a binarized result.
3. Statistical Result Aggregation: The binarized results from the previous step are aggregated for all samples, with a 2-sample t-statistic used to provide a singular p-value on the range of 0 (D_q was not included in training) to 1 (D_q was included in training).

3 Systematic Evaluation of EMA's Robustness

We systematically evaluate EMA's robustness against various techniques known to impair membership inference potency. We detail our experimental setup in Section 3.1 and present results in Section 3.2. We then conduct ablation studies to understand EMA's poorer performance on COVIDx in Section 3.3.

3.1 Experimental setup

Datasets and models We conducted experiments using three datasets with different model architectures, including MNIST (LeCun et al., 1998) with MLP, COVIDx (Wang et al., 2020) with ResNet-18 (He et al., 2016), and the Location dataset (Yang et al., 2016) with MLP. More information about these datasets and their associated models is available in Appendix A.

Adversarial Methods. We evaluated EMA's robustness against techniques known to impair membership inference performance, including dropout, pruning, and MemGuard (Jia et al., 2019):

- Dropout (Srivastava et al., 2014) disables nodes at probability p during training to learn more generalizable patterns (Galinkin, 2021; Salem et al., 2018). We vary probabilities for dropout from 0 to 0.9 in increments of 0.1.
- Unstructured global magnitude pruning (Brownlee, 2021) is also a countermeasure for membership inference (Wang et al., 2021). Similar to dropout probability, we vary the sparsity—the fraction of zeroed-out weights—between 0 and 1 in increments of 0.1 to identify performance cliffs.
- MemGuard (Jia et al., 2019) adds adversarial noise to hinder membership inference. We found a bug in their code; therefore, we present results with both buggy and corrected implementations.

Metrics. For each adversarial method, we consider the potential impact it might have on both the effectiveness of EMA and model utility. Therefore, we report the drop in both EMA efficacy (i.e., "the EMA cost") and the drop in the target model accuracy (i.e., "the classification cost"). The EMA accuracy is averaged across 378 auditing runs of various configurations; See Appendix A for details.

3.2 Results

Baseline performance. Table 1 summarizes the baselines for MNIST, COVIDx, and Location, with the results of (Huang et al., 2022) being replicated on MNIST and COVIDx. Note that the 51.50% test accuracy of Location is consistent with previous work (Jia et al., 2019).

	Classes	Test Acc.	EMA Acc.
MNIST + MLP	10	96.24%	92.03%
COVIDx + ResNet-18	3	86.50%	92.13%
Location + MLP	30	51.50%	98.24%

Table 1: Target model accuracies and EMA accuracies.

Dropout's impact on EMA can be reasonably mitigated by careful metric selection. Figure 1 shows the effects of dropout on both EMA accuracy and target model accuracy. EMA is robust to dropout for MNIST and Location, with poor auditing performance only occurring at unreasonably high classification costs. On COVIDx, however, dropout impairs EMA performance by 46% while *increasing* model accuracy. Such results suggest that dropout can act as a practical hindrance to EMA.

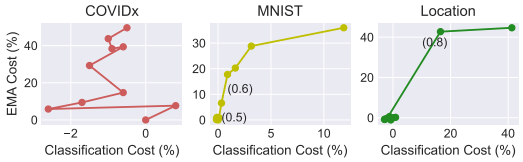

Figure 1: The reduction caused by dropout in EMA accuracy (i.e., EMA cost) v.s. the reduction in model's accuracy (i.e., Classification cost). Dropout values for key datapoints are annotated, *e.g.* (0.5).

To identify the shortcomings of EMA on dropped-out models, the metrics used in the auditing algorithm were fine tuned by a novel metric-set analysis. In this process, audits were performed using each metric individually and in various subsets. Surprisingly, we found that the negative entropy metric *worsens* auditing performance on all datasets and dropout levels, including the baselines. Figure 2 shows the improved auditing performance on COVIDx by removing negative entropy from the combined metric set.

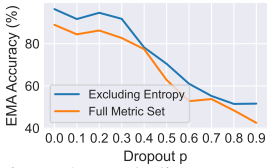

Figure 2: Excluding entropy boosts EMA performance.

EMA is not robust against pruning. As seen from Figure 3, the Location model is especially robust to pruning, as we incur an EMA loss only after a classification loss of over 10% (at a sparsity of 77.5%). However, MNIST displays an immediate dropoff in EMA accuracy even at the 10% sparsity level. COVIDx is yet more staggering: at the 20% sparsity level, average EMA accuracy drops from 92.13% to 69.28% with only a 2.5% classification loss. Metric-set analysis was also performed on the pruned models. Unlike the dropout case, where removing the negative entropy metric significantly improved performance, we found that the poor EMA accuracy on pruned COVIDx models resulted from a low efficacy of **all** metrics when a sparsity of over 20% was applied. Of particular note is the decrease in potency of the "correctness" metric, which drops from 92.58% to 67.58% at 20% sparsity.

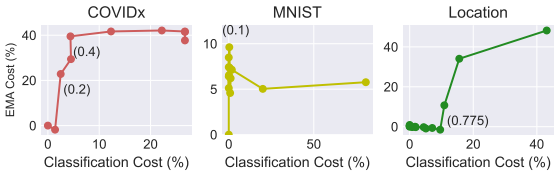

Figure 3: EMA cost v.s. Classification cost under pruning. Sparsity values for key datapoints are annotated, *e.g.* (0.775).

MemGuard is a comparatively less practical adversarial technique. We noticed a bug in MemGuard's original implementation, where the noisy logits were not mixed with their denoised counterparts. Results are thus presented with both mixed (correct) and unmixed (buggy) logits. As seen from Table 2, MemGuard has a severe effect on the potency of EMA. However, such results come with an overly-high classification loss—dropping accuracy to 37.5% for the mixup case and 21.0% without mixup. Another limitation of MemGuard is that the auditor cannot, in theory, query the target model directly. Instead of altering the model itself, as dropout does, MemGuard simply serves a different set of output logits to the end user. Such a fact, as well as the low classification accuracy MemGuard induces, limits the practicality of its defense against model auditing.

	EMA Acc.	**Test Acc.**
Baseline	95.58%	51.50%
MemGuard (buggy)	63.82%	21.00%
MemGuard (correct)	52.34%	37.50%

Table 2: Effects of MemGuard on Location.

3.3 Discussion

Given the low robustness of EMA to dropped-out and pruned COVIDx models, we examine the effects of model size and overfitting in order to understand EMA performance. In particular, the COVIDx model is more generalized and has more parameters than the MNIST or Location models.

Model generalization indicates EMA performance. The COVIDx results in Section 3.2 and the ablations in Appendix C support the hypothesis that *generalized models* pose a challenge to Ensembled Membership Auditing. Such a hypothesis explains why regularization techniques like dropout and pruning yield a significant decrease in EMA performance, albeit at the cost of target model accuracy.

False Negative Error Pattern. Figure 4 shows EMA accuracy versus dataset size on a pruned MNIST model. In contrast to the baseline models, where EMA produces false positive labels on small dataset sizes, we find that EMA is also *under-sensitive* when dropout is applied. Specifically, EMA produces false negatives (non-membership) on larger query datasets, which yields a significant performance decrease. This False Negative Error Pattern (FNEP) was not witnessed by (Huang et al., 2022).

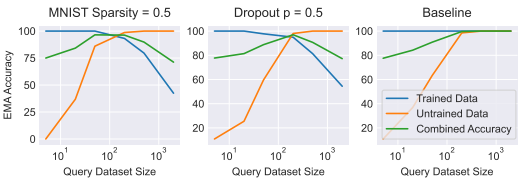

Figure 4: Examining the trends in EMA Accuracy by dataset size for sparse, dropped-out, and baseline MNIST models. Regularization (in the left two graphs) induces false negatives at large dataset sizes, while false positives are ever-present for smaller dataset sizes.

4 EMA-Zero: Towards Calibration-free Unlearning Verification

Method. EMA faces a significant constraint in that it relies on an external dataset for calibration model training (Step 1 in Section 2). Consider a scenario within Federated Learning, where a hospital cannot share a separate dataset with an external auditor due to concerns about compromising patient Protected Health Information (PHI). To provide similarly-distributed data without compromising confidentiality, we propose a solution: the provider can train a cost-effective Generative Adversarial Network (GAN) Goodfellow et al. (2014) on their dataset and offer the auditor access to a black-box generator. The auditor can then leverage this GAN to generate synthetic images as needed, with the target model providing labels. These synthetic data and labels can be employed to train the calibration model and to obtain thresholds for each metric, resulting in a novel pipeline termed "EMA-Zero." We provide a flow chart for this process in Appendix B.

Results. We use a simple 3-layer, 256-unit MLP for both the generator and discriminator. To determine the importance of calibration data quality on auditing performance, we vary the amount of data used to train the GAN, from 100 samples to the full dataset. We present results on COVIDx below, with consistent results on MNIST in Appendix C. As shown in Table 3, EMA efficacy seems to be weakly correlated with the quality of the GAN generating the calibration dataset. Such results, across both MNIST and COVIDx, yielded the following key conclusions:

1. **(Lack of) Importance of Calibration Dataset:** Figure 8 and Table 3 show that utilizing datasets generated by a low-quality GAN can yield similar EMA efficacy as using real datasets. One explanation is that calibration in EMA is effective not because the calibration dataset mimics the training dataset, but because of the general process of identifying discrepancies between the calibration model's training and testing results.
2. **Efficacy of EMA-Zero:** This work was able to provide a novel adaptation of EMA for the scenario with no available calibration dataset. The EMA-Zero pipeline was able to approach baseline EMA performance with as few as 100 training samples, demonstrating its practical applicability.

Training Size	EMA Acc.
0 (Pure Noise)	84.43
100	84.48
200	83.16
1000	87.38
2000	85.35
4000	84.10
Zero noise	86.17

Table 3: EMA Accuracy (%) v.s. GAN training size. The last row corresponds to EMA-Zero with a GAN trained on 4000 unnoised images.

5 Conclusions

This study examined the efficacy of Ensembled Membership Auditing (EMA). We first evaluated EMA's robustness and found that pruning and dropout significantly impair EMA efficacy while maintaining classification accuracy. We then introduced a novel metric-set analysis and found that removing negative entropy improved EMA on all models. Ablation studies further indicated that EMA was less robust with larger and convolutional models. Orthogonally, we proposed a novel pipeline, EMA-Zero, for situations where the auditor cannot access an in-distribution calibration dataset. Even with data generated from a low-resource GAN, EMA-Zero was able to reach baseline EMA levels for both MNIST and COVIDx. Such results demonstrate EMA-Zero's applicability in novel collaborative learning settings, where auditors can operate effectively without being exposed to confidential information.

Future work will develop theoretical guarantees for EMA, mitigate false negatives induced by regularizers, expand the applicability of our approach to generative models, and extend the study to explore alternative data auditing methods.

References

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. (2021). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

Brownlee, J. (2021). Gentle introduction to vector norms in machine learning.

Galinkin, E. (2021). The influence of dropout on membership inference in differentially private models. *arXiv preprint arXiv:2103.09008*.

Gao, X., Ma, X., Wang, J., Sun, Y., Li, B., Ji, S., Cheng, P., and Chen, J. (2022). Verifi: Towards verifiable federated unlearning. *arXiv preprint arXiv:2205.12709*.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. (2019). Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. (2013). Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*.

Gupta, V., Jung, C., Neel, S., Roth, A., Sharifi-Malvajerdi, S., and Waites, C. (2021). Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, Y., Huang, C.-Y., Li, X., and Li, K. (2022). A dataset auditing method for collaboratively trained machine learning models. *IEEE Transactions on Medical Imaging*.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. (2021). Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pages 2008–2016. PMLR.

Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. (2019). Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 259–274.

Kermany, D., Zhang, K., Goldbaum, M., et al. (2018). Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2):651.

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

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., et al. (1995). Comparison of learning algorithms for handwritten digit recognition. In *International conference on artificial neural networks*, volume 60, pages 53–60. Perth, Australia.

Legislature, C. S. (2018). California consumer privacy act.

Liu, G., Ma, X., Yang, Y., Wang, C., and Liu, J. (2020). Federated unlearning. *arXiv preprint arXiv:2012.13891*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. (2018). Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. (2021). Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. (2023). Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Wang, L., Lin, Z. Q., and Wong, A. (2020). Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549.

Wang, Y., Wang, C., Wang, Z., Zhou, S., Liu, H., Bi, J., Ding, C., and Rajasekaran, S. (2021). Against membership inference attack: Pruning is all you need.

Weng, J., Yao, S., Du, Y., Huang, J., Weng, J., and Wang, C. (2022). Proof of unlearning: Definitions and instantiation. *arXiv preprint arXiv:2210.11334*.

Wu, Y., Dobriban, E., and Davidson, S. (2020). Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, pages 10355–10366. PMLR.

Yang, D., Zhang, D., and Qu, B. (2016). Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):1–23.

Ye, J., Fu, Y., Song, J., Yang, X., Liu, S., Jin, X., Song, M., and Wang, X. (2022). Learning with recoverable forgetting. In *European Conference on Computer Vision*, pages 87–103. Springer.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Ying, Z., Zhang, Y., and Liu, X. (2020). Privacy-preserving in defending against membership inference attacks. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 61–63.

Zhou, J., Li, H., Liao, X., Zhang, B., He, W., Li, Z., Zhou, L., and Gao, X. (2023). Audit to forget: A unified method to revoke patients' private data in intelligent healthcare. *bioRxiv*, pages 2023–02.

A Experimental Details

Datasets and models. MNIST LeCun et al. (1998) is a computer vision benchmark with a classification task of 10 handwritten digits. A subset of 10,000 images were utilized for training the target 3-layer, 256-node multilayer perceptron (MLP). The OOD fold was collected from the Street View House Numbers (SVHN) dataset Goodfellow et al. (2013).

COVIDx Wang et al. (2020) dataset contains Chest X-Ray (CXR) images of COVID-positive, pneumonic, and healthy patients. In contrast to the fully-connected perceptron used for MNIST, a ResNet18 model was used for COVIDx trained on 4,000 images. A separate OOD fold was also collected from the Child-XRay dataset Kermany et al. (2018).

Location dataset Yang et al. (2016) is a tabular benchmark that has been widely used in MIA literature Shokri et al. (2017); Yeom et al. (2018). Location has been used widely in previous works relating to membership inference Ying et al. (2020). In contrast to the image-based datasets, Location contains tabular data of 5,010 samples with 446 binary features and a 30-class prediction task Shokri et al. (2017).

Calculation of EMA accuracy. We vary the configurations among the calibration model (6 choices) and query datasets (7 different data folds and 9 choices of sizes) as below to obtain $6 \times 7 \times 9 = 378$ EMA results for each (dataset, model) benchmark:

1. Calibration Model Training: Six calibration models, each containing the same architecture as the target model, are trained on a held-out dataset from the same distribution as the training data. Across the six models, the percentage of noise applied to the calibration dataset is varied by parameter k, where $k = 20$ has 20% of its data samples perturbed by either noise or rotation. The models are thus $C_0, C_{10}, C_{20}, C_{30}, C_{40}, C_{50}$.
2. Query dataset: D_q consists of five data folds from the target model's training set, one untrained fold from the same data distribution, and one fold from a separate out-of-distribution (OOD) dataset. For each D_q, we run the results using nine different sizes of truncates.

For each run, we obtain a p-value and convert it to an accuracy by identifying the percent change from the ground truth p-value (0 or 1). All such percentages across the 378 runs are averaged, with equal weight given to the training folds, testing fold, and OOD fold.

B EMA-Zero Pipeline

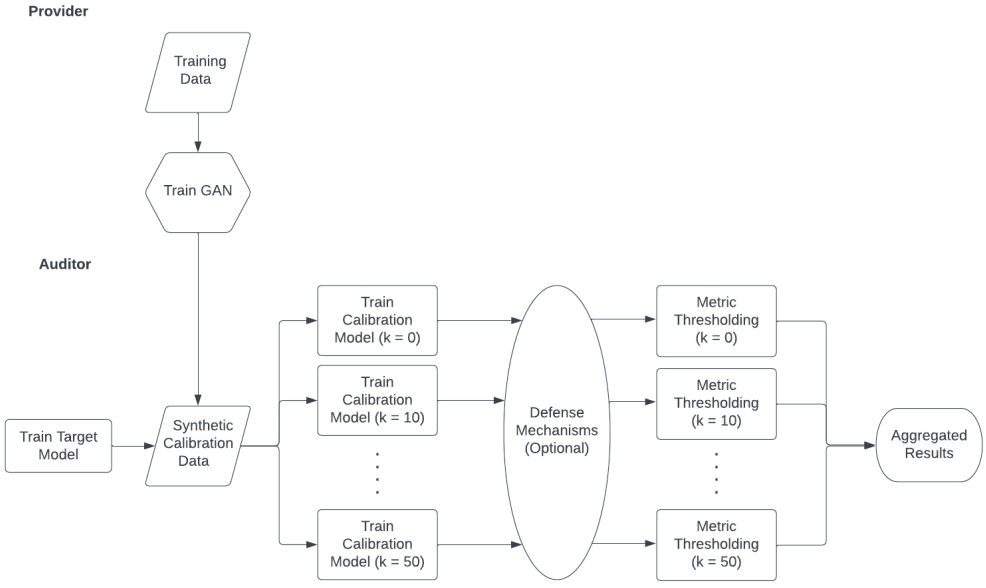

Figure 5: EMA-Zero: Ensembled Membership Auditing with synthetic data.

EMA-Zero supplants real calibration data for synthetic data produced by an externally-trained GAN. The data is then passed down to the original EMA processes of calibration model training, thresholding, and result aggregation.

C Further Results

Ablation: Convolutional vs. Fully-Connected Models. We replaced the MNIST MLP with the convolutional LeNet5 model to isolate the effects of model type LeCun et al. (1995).

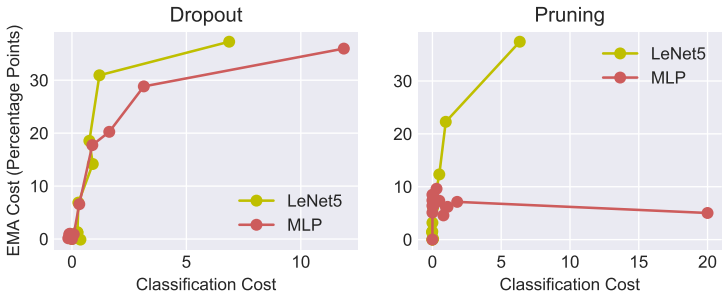

Figure 6: Loss in average EMA accuracy vs. loss in classification accuracy for LeNet5 and MLP on MNIST.

As seen from Figure 6, EMA is less robust to the convolutional LeNet5 model than it is to the fully-connected MLP. Specifically, at 2% classification loss, we see that EMA performs up to 10% more poorly on the convolutional model with dropout and up to 25% more poorly with pruning. Such results substantiate the COVIDx findings that convolutional models may pose a greater challenge to EMA than fully-connected ones. An intuitive explanation is that convolutional models capture more generalized patterns, making them harder to audit.

Ablation: Model Size. We also replaced the original 3-layer MNIST MLP with a 2-layer model to examine the effects of model size.

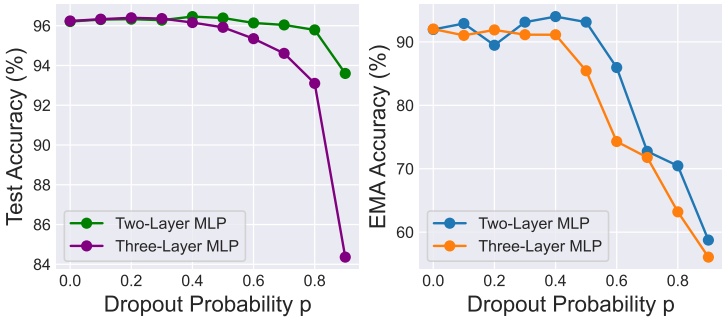

Figure 7: EMA Accuracy and classification accuracy versus dropout for the small and large MLP models.

EMA's baseline performance on the two models was similar, achieving 92.03% on the three-layer and 91.96% on the two-layer. As shown in Figure 7, however, both auditing and classification performance fared better on the smaller model when dropout was induced, with the difference in EMA accuracy for the two models up to 15% (at a dropout level of 0.5).

EMA-Zero with MNIST. The strong performance of EMA-Zero on COVIDx was replicated on MNIST. Figure 8 shows EMA-Zero accuracy using the LeNet5 model from Section 3.3.

Interestingly, the quality of the GAN-generated data does not significantly affect EMA performance. Excluding the GAN experiment with a training size of 20, EMA accuracy seems to hover around 90% regardless of the strength of the synthetic calibration data, close to the real-data baseline of 91.56%. It is further apparent that the entirety of the error arises from the false positive error pattern noted by Huang et. al. 2022 with none of the false negative error (FNEP) described in Section 3.2.

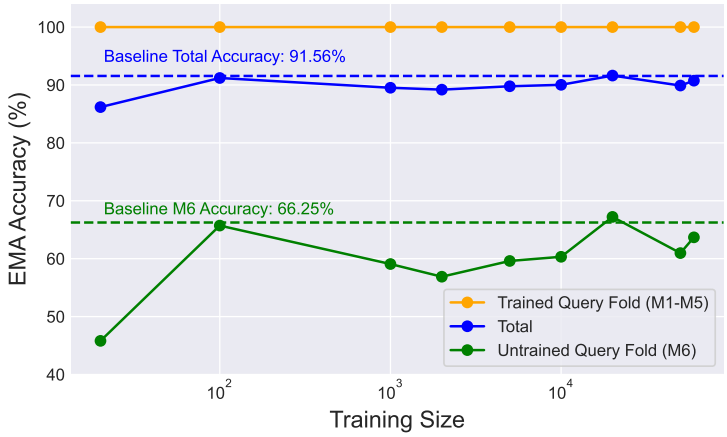

Figure 8: Average EMA Accuracy with Calibration Models trained on GAN-generated datasets, with he number of training samples for the GAN varied. EMA accuracy is broken down into the trained query data folds (M1-M5) and untrained query fold (M6).