A Dataset Auditing Method for Collaboratively Trained Machine Learning Models

Yangsibo Huang, Chun-Yin Huang, Xiaoxiao Li *Member, IEEE*, and Kai Li *Fellow, IEEE*

***Abstract*— Dataset auditing for machine learning (ML) models is a method to evaluate if a given dataset is used in training a model. In a Federated Learning setting where multiple institutions collaboratively train a model with their decentralized private datasets, dataset auditing can facilitate the enforcement of regulations, which provide rules for preserving privacy, but also allow users to revoke authorizations and remove their data from collaboratively trained models.**

This paper first proposes a set of requirements for a practical dataset auditing method, and then present a novel dataset auditing method called Ensembled Membership Auditing (EMA). Its key idea is to leverage previously proposed Membership Inference Attack methods and to aggregate data-wise membership scores using statistic testing to audit a dataset for a ML model. We have experimentally evaluated the proposed approach with benchmark datasets, as well as 4 X-ray datasets (CBIS-DDSM, COVIDx, ChildXRay, and CXR-NIH) and 3 dermatology datasets (DERM7pt, HAM10000, and PAD-UFES-20). Our results show that EMA meet the requirements substantially better than the previous state-of-the-art method.

Our code is at: https://github.com/Hazelsuko07/EMA.

***Index Terms*— Privacy, Dataset Auditing, Medical Image Classification**

I. Introduction

Advances in artificial intelligence (AI) have led to many disruptive innovations in science and technology. However, AI for healthcare is hindered by restricted accesses to decentralized data silos due to privacy concerns and regulations such as Health Insurance Portability and Accountability Act (HIPAA) [1] that prohibit the sharing of sensitive healthcare data.

Federated learning (FL) [2], [3] is a promising paradigm to mitigate such concerns as it allows multiple participants to collaboratively train a model without transferring decentralized datasets to a central server. However, recent regulations such as California Consumer Privacy Act (CCPA) [4] and General Data Protection Regulation (GDPR) [5] give individuals more control over their data, such as revoking previous authorizations and removing data from collaboratively trained models.

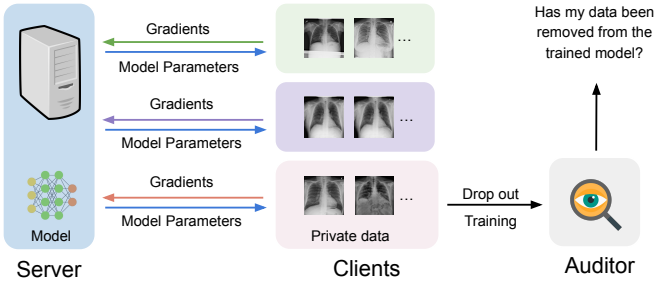

Fig. 1: In a Federated Learning setting where multiple institutions (clients) collaboratively train a model, one institution withdraws and later uses dataset auditing to ensure its dataset is removed from the model.

Dataset auditing for machine learning models is a verification process to facilitate the implementation of such regulations. If an institution withdraws its participation from an FL agreement and requires removing its dataset from a collaboratively trained model, dataset auditing can be used to ensure that its dataset is not used in training the latest model, as shown in Figure 1.

In addition, dataset auditing can be used to check if a suspicious dataset or poisoned dataset is used in training a model, avoiding the potential hazards of using malicious datasets. Dataset auditing can also be used to monitor whether a machine learning model meets its requirement. For example, when the FDA examines whether an AI model is trained on the given collection of datasets that include biased group data for AI fairness [7], dataset auditing can be used to ensure that datasets including underrepresented groups are used in training a model.

In this paper, we first propose a set of design requirements for a practical dataset auditing method. A practical method should have minimal input requirements, be time-efficient for large audited dataset and model, be robust with audited datasets of different scales and distributions, and ideally provide certain theoretical guarantees. As dataset auditing is an under-explored topic, no previous method meets all requirements well.

We then present an Ensembled Membership Auditing (EMA) method inspired by previous work on membership inference attacks [8]on a trained model (see Fig. 2). EMA

Y. Huang and C.Y. Huang contribute to this work equally.
Y. Huang and K. Li are with Electrical and Computer Engineering/Computer Science Department, Princeton University, NJ 08540, USA
C.Y. Huang and X. Li is with Electrical and Computer Engineering Department, the University of British Columbia, Vancouver, BC V6T 1Z4 Canada.
Correspondence to: Y. Huang (yangsibo@princeton.edu), X. Li (xiaoxiao.li@ece.ubc.ca)

	Ideal requirements	CaliForget [6]	EMA (ours)	**Reference**
Inputs	Access to audited dataset D_a	Yes	Yes	
	Black-box access to audited model	Yes	Yes	Section III-A
		Calibration dataset D_{cal}	D_{cal}	Section III-A
Efficiency	Time and space efficient	Train a model with D_{cal}, and a model with D_a	Train a model with D_{cal}	Section III-A
Robustness	Allow various distributions of D_a	Not stable in several cases	Stable	Section VI-C.1
	Allow small $\|D_a\|$	Work when $\|D_a\| \geq 800$	Work when $\|D_a\| \geq 50$	Section VI-C.2
		Sensitive to the quality of D_{cal}	Not so sensitive	Section VI-C.3
Analysis	Theoretical guarantee	No	No	Future work

TABLE I: Requirements for dataset auditing. These requirements can be viewed as a checklist of evaluating objectives to optimize a dataset auditing algorithm. We categorize them into: inputs, efficiency, robustness, and analysis. CaliForget and EMA meet some requirements well. (See Section IV for details).

consists of a 2-step procedure: infers whether the model memorizes each sample of the audited dataset based on *multiple membership metrics*, and ensembles these membership metrics and utilizes statistical tools to aggregate the sample-wise results and obtain a final auditing score. We show that this method meets the design requirements better than the previous state-of-the-art [6].

Our contributions are summarized as follows:

1) **Requirements for dataset auditing.** We propose a set of requirements for a practical dataset auditing method in terms of *inputs*, *efficiency*, *robustness* and *theoretical guarantee*.
2) **Ensembled Membership Auditing** (EMA)**.** We propose a novel method for dataset auditing. We demonstrate that EMA meets the requirements better than the previous state-of-the-art CaliForget [6].
3) **Experimental evaluations with medical datasets.** Our evaluation includes three types of datasets, including non-medical benchmark datasets (MNIST and SVHN), Chest X-ray datasets (COVIDx, Child-XRay, CXR-NIH and CBIS-DDSM), and skin lesion datasets (HAM and PAD-UFES-20). Our experiments demonstrate that EMA approach is *robust* under various practical settings, including the conditions that the previous method fails.

A preliminary version of this work was done, known as [9]. This manuscript extends the preliminary work in two ways. First, we propose the requirements for dataset auditing. Second, we have conducted extensive experiments on new datasets (e.g. dermatology datasets such as HAM10000 and PAD-UFES-20) and detailed ablation studies to validate the generalizability of EMA under practical conditions and settings.

II. Related Work

Dataset auditing for machine learning models is an under-explored area of research. The only previous work on dataset auditing for machine learning models is CaliForget [6], an algorithm to verify if an audited dataset is used by an audited machine learning model, with the help of a calibration dataset. The method is based on Kolmogorov-Smirnov (KS) distance and requires training a model on the audited dataset. However, the method fails under certain practical conditions, such as when the audited dataset is similar to the training dataset or when the calibration dataset is not of high quality.

One of the components of our proposed method is membership inference attack. ML models such as DNNs are often overparameterized, which means that they have sufficient capacity to memorize information about their training datasets [10]–[12]. Such ML models are vulnerable to membership inference attacks (MIAs) [8], [13]–[15], which aim to infer whether a single data record was used to train an ML model or not (see Section V-A for a formal problem formulation).

Dataset auditing in general is an important process to facility the implementation of data protection regulations. Health Insurance Portability and Accountability Act (HIPAA) [1] enacted in 1996 is a federal regulation to protect sensitive patient health information from being disclosed without patients' consent. More recent regulations such as California Consumer Privacy Act (CCPA) [4] and General Data Protection Regulation (GDPR) [5] give individuals control over their data including the right to know how data are used and shared, the right to delete, the right to opt-out, and the right to exercise their rights without discrimination. These regulations require companies or institutions to remove data from their systems and models upon their requests.

Machine unlearning introduced by [16] studies how to remove data from ML models without retraining. A large body of recent work focus on different models of unlearning [17]–[23].The goal of dataset auditing for ML models is different from that of machine unlearning; it aims at verifying if a given dataset is used in training a machine learning model.

Another related area is to quantify properties of machine learning models. By quantifying privacy leakage or bias of machine learning models such as attacks for inferring training data points, one can show the vulnerability of machine learning models [11], [13], [24], [25]. There are also studies to quantify the societal impact of a machine learning model [26], [27], especially the biases that ML models may induce [28]. They develop toolkits that allow users to test machine learning models with regards to bias and fairness metrics for different population subgroups.

III. Preliminary

A. Problem formulation

The goal of dataset auditing for a machine learning model is to answer the question if an audited dataset is used in training the machine learning model. In order to meet requirements in

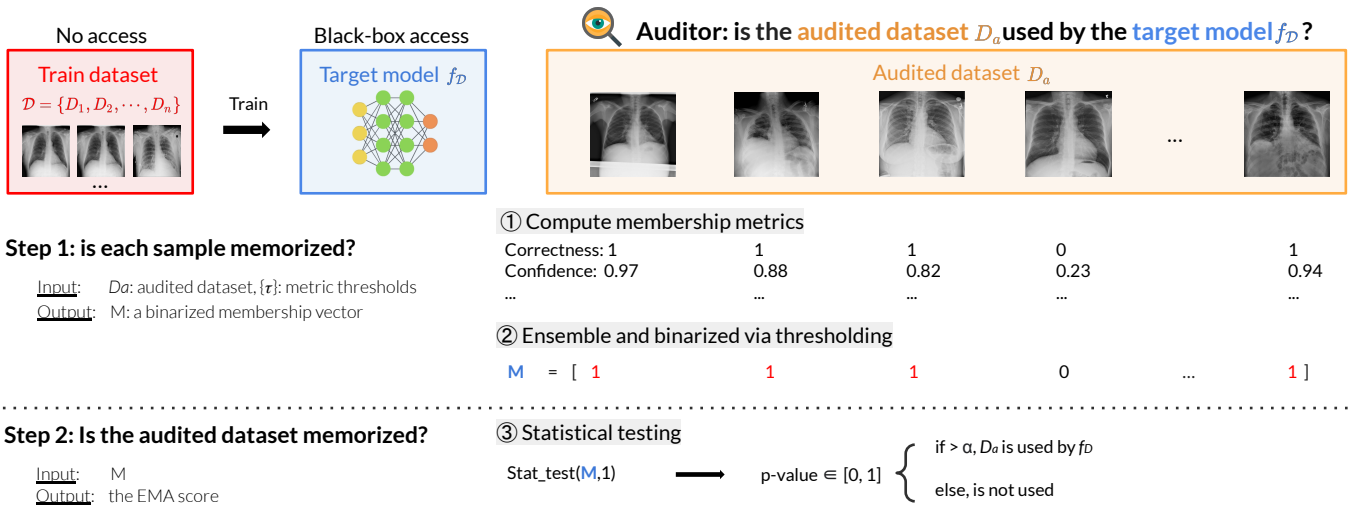

Fig. 2: Illustration of our proposed EMA method. The auditor has black-box access to the target model (blue box) and no access to its training dataset (red box), and aims to verify if a audited dataset from an institution has been used by the target model. EMA consists of two steps: 1) the auditor first infers if each sample in the audited dataset is used by the target model; 2) then it ensembles the results and see if the whole audited dataset is used. See Algorithm 1 for more details.

Table I, we assume that an auditor has access to the model's outputs but not model parameters. This is usually referred to as a "black-box" setting.

Our formulation of the data auditing problem is as follows: suppose $\mathcal{D}$, a collection of n training datasets $\{D_1, D_2, \cdots, D_n\}$, is sampled from a given distribution $\mathbb{D} \subset \mathbb{R}^d$, where d denotes the input dimension. A machine learning model $f_{\mathcal{D}} : \mathbb{R}^d \to \mathcal{C}$ is trained on $\mathcal{D}$ to learn the mapping from an input to a label in the output space $\mathcal{C}$. We denote the inference with a data point $x \in \mathbb{R}^d$ as $f_{\mathcal{D}}(x)$. The auditor with black-box access to the model $f_{\mathcal{D}}$ aims to tell if an audited dataset D_{a} has been used to train $f_{\mathcal{D}}$. Specifically, the auditor has access to:

- The architecture of $f_{\mathcal{D}}$ and the algorithm to train it;
- The audited dataset D_{a};
- $f_{\mathcal{D}}(D_{\mathrm{a}})$, probability outputs of the audited data D_{a} on the model $f_{\mathcal{D}}$;
- The distribution of the original training data points.

Note that the auditor does **not** have access to the original training dataset collection $\mathcal{D}$ except D_{a}, nor the network parameters of $f_{\mathcal{D}}$.

B. Previous Method

Apart from the audited dataset D_{a}, the previous approach CaliForget [6] requires using a calibration dataset D_{cal} to run the auditing procedure. Specifically, it runs the following steps in sequence:

1) Train $f_{D_{\mathrm{cal}}}$ using D_{cal}, with the model architecture and training algorithm as the target model $f_{\mathcal{D}}$;
2) Train $f_{D_{\mathrm{a}}}$ using D_{a}, with the model architecture and training algorithm as the target model $f_{\mathcal{D}}$;
3) Compute the following criteria:

$$\rho_{\mathrm{CF}} = \frac{\mathrm{KS}(f_{\mathcal{D}}(D_{\mathrm{a}}), f_{D_{\mathrm{a}}}(D_{\mathrm{a}}))}{\mathrm{KS}(f_{D_{\mathrm{cal}}}(D_{\mathrm{a}}), f_{D_{\mathrm{a}}}(D_{\mathrm{a}}))}, \tag{1}$$

where $\mathrm{KS}(\cdot)$ is the Kolmogorov-Smirnov (K-S) distance between two distributions.

CaliForget interprets ρ_{CF} such that $\rho_{\mathrm{CF}} < 1$ indicates the audited dataset collection D_{a} has been used by $f_{\mathcal{D}}$. The motivation is that if $\mathcal{D}$ contains samples from D_{a}, the K-S distance between $(f_{\mathcal{D}}(D_{\mathrm{a}})$ and $f_{D_{\mathrm{a}}}(D_{\mathrm{a}}))$ will be very small, and conceivably is much smaller than the reference distance calculated based on the calibration dataset $KS(f_{D_{\mathrm{cal}}}(D_{\mathrm{a}}), f_{D_{\mathrm{a}}}(D_{\mathrm{a}}))$ (i.e. $\rho_{\mathrm{CF}} < 1$). On the contrary, if $\mathcal{D}$ has no samples from D_{a}, then the value of $KS(f_{\mathcal{D}}(D_{\mathrm{a}}), f_{D_{\mathrm{a}}}(D_{\mathrm{a}}))$ depends on the statistical overlap of $\mathcal{D}$ and D_{a}, which can be calibrated by referring to a reference distance $KS(f_{D_{\mathrm{cal}}}(D_{\mathrm{a}}), f_{D_{\mathrm{a}}}(D_{\mathrm{a}}))$.

However, by construction, the ρ_{CF} criteria may fail under the following conditions:

- when the audited dataset D_{a} is very similar to the original training dataset collection $\mathcal{D}$, the numerator is small, which will lead to a false-positive result (i.e. predict 'not used' as 'used');
- when the calibration dataset D_{cal} is of low quality, the denominator is small, which will lead to a false-negative result (i.e. predict 'used' as 'not used').

Section VI provides experimental results of the above limitations of using ρ_{CF}.

IV. Requirements For Dataset Auditing

To address the particle needs of data auditing in healthcare applications, we introduce a set of requirements for designing a dataset auditing system. These requirements can be viewed as a checklist of evaluating objectives to optimize a dataset auditing algorithm. We categorize them into: *inputs, efficiency, robustness*, and *analysis*, as summarized in Table I.

"Inputs" are what a dataset auditing system requires to run a dataset auditing algorithm. The ideal and minimal requirements are accessing to the audited dataset D_{a} and the audited machine learning model $f_{\mathcal{D}}$. For instance, if the

audited model is deployed as a set of machine learning APIs (i.e., a black-box model), the auditing algorithm should not assume knowledge of the model's private parameters such as weights. Ideally, the auditing algorithm does not require any other information.

"Efficiency" is about how efficient the algorithm is in terms of time and space requirements when auditing a large auditing dataset for a large machine learning model.

"Robustness" requires that the auditing algorithm is robust with different audited datasets, regardless of their distributions and sizes. If the auditing algorithm requires using a calibration dataset, the algorithm should be robust with different qualities.

"Analysis" says that it is important to have a theoretical guarantee for auditing success.

The previous work CaliForget and our approach do not meet two requirements well. First, both CaliForget and EMA require a calibration dataset which has similar distribution to the dataset collection $\mathcal{D}$. Second, neither CaliForget nor EMA has a theoretical guarantee for auditing success. It is important for future work to address these shortcomings.

V. The Proposed Method

We propose Ensembled Membership Auditing (EMA), a two-step procedure to audit a dataset and a trained model: sample-wise membership inference and membership statistics ensemble (see Algorithm 1) that better fit the requirements in Sec IV than the existing methods. The following describes each step in detail.

A. Sample-wise Membership Inference

This step infers whether the model memorizes each sample of the audited dataset based on multiple membership metrics. Our inference method builds on the membership inference attack (MIA) in the black-box setting [8], where the auditor observes the outputs of the auditing machine learning model with given inputs without knowing more details such as model parameters.

We also use a calibration dataset, which has distributions similar to the training dataset, to establish thresholds for several metrics as decision rules. The idea to use multiple metrics to define decision rules is originally proposed by [15] and shown to achieve better performance than using a machine learning model trained on the calibration data ([8], [13], [29]).

Formally, given the target model f_D, which is trained with training dataset D, and an audited dataset $D_{\rm a}$, the first step infers if each sample in $D_{\rm a}$ is used by f_D (see Algorithm 1, line 2 to line 6). The auditor first computes $\tau_1, \cdots, \tau_m$, thresholds for m different metrics by running a standard membership inference pipeline [15] on the calibration set. These 3 metrics are:

- Correctness: $g_{\rm corr}(f,(x,y)) = \mathbf{1}\{\arg\max_i \ f(x)_i = y\}$[1]
- Confidence: $g_{\rm conf}(f,(x,y)) = f(x)_y$[2]
- Negative entropy: $g_{\rm entr}(f,(x,y)) = \sum_i f(x)_i \log(f(x)_i)$[3]

[1]Leino et al. [30] suggest a trained ML model is more likely to give correct prediction on training data than on test data.

[2]Yeom et al. [14] show that training data has a higher confidence in predicting the correct label than test data.

[3]Shokri et al. [8] show that training data usually have lower prediction entropy (i.e. higher negative entropy) than test data.

Algorithm 1 Ensembled Membership Auditing (EMA)

Input: A, the training algorithm; f_D, the target model; $D_{\rm a}$, the audited dataset; $D_{\rm cal}$, the calibration dataset; $g_1, \cdots, g_m$, m different metrics for membership testing.

Output: $\rho_{\rm EMA} \in [0,1]$, the possibility that $D_{\rm a}$ is used by f_D

1: **procedure** EnsembledMembershipAuditing
2: $\tau_1, \cdots, \tau_m \leftarrow$ InferThres$(A, D_{\rm cal}, g_1, \cdots, g_m)$ ▷ See Algorithm 2
3: $\mathbf{M} \leftarrow \mathbf{0}$ ▷ $\mathbf{M} \in \{0,1\}^{|D_{\rm a}|}$, the inferred membership of each sample in $D_{\rm a}$
4: **for** $(x_i, y_i) \in D_{\rm a}$ **do**
5: $\mathbf{M}_i \leftarrow \mathbf{1}\{g_1(f_D,(x_i,y_i)) \geq \tau_1\} \cup \mathbf{1}\{g_2(f_D,(x_i,y_i)) \geq \tau_2\} \cup \cdots \cup \mathbf{1}\{g_m(f_D,(x_i,y_i)) \geq \tau_m\}$
6: **end for**
7: $\rho_{\rm EMA} \leftarrow$ pvalue$(\mathbf{M}, \mathbf{1})$ ▷ pvalue() returns the p-value of a two-sample statistical test, which determines if two populations are from the same distribution
8: **return** $\rho_{\rm EMA}$
9: **end procedure**

To select thresholds to identify training data, we define balanced accuracy on calibration data based on the balanced accuracy regarding True Positive Rate (TPR) and True Negative Rate (TNR):

$$\mathrm{BA}(\tau) = \frac{TPR(\tau) + TNR(\tau)}{2} \tag{2}$$

where given a threshold τ, $TPR(\tau) = \sum_{s \in D_{\rm cal}^{\rm train}} \mathbf{1}\{g_i(s) \geq \tau\} / |D_{\rm cal}^{\rm train}|$, and $TNR(\tau) = \sum_{s \in D_{\rm cal}^{\rm test}} \mathbf{1}\{g_i(s) \geq \tau\} / |D_{\rm cal}^{\rm test}|$.

The best threshold is selected to maximize the balanced accuracy (see Algorithm 2). For each sample in $D_{\rm a}$, it will be inferred as a member or used by the target model, if it gets a membership score higher than the threshold for at least one metric (Algorithm 1, line 3 to 6). The auditor stores the membership results in $\mathbf{M} \in \{0,1\}^{|D_{\rm a}|}$: $\mathbf{M}_i = 1$ indicates that the i-th sample in $D_{\rm a}$ is inferred as used by f_D, and $\mathbf{M}_i = 0$ indicates otherwise.

B. Membership Statistics Ensemble

This step ensembles multiple metrics of the sample-wise membership inference step to obtain a final auditing score (see Algorithm 1).

Given $\mathbf{M}$, the sample-wise inference results from step 1, the auditor infers if the whole audited dataset is used. A simple approach is to perform majority voting on $\mathbf{M}$, however, the state-of-the-art MIA approaches [15] achieve only ∼70% accuracy with benchmark datasets. Majority voting may not achieve reliable results.

The unreliability of a single entry in $\mathbf{M}$ motivates us to consider using the distribution of $\mathbf{M}$: ideally, if an audited dataset D_a^* is used, it should give $\mathbf{M}_{D_a^*} = \mathbf{1}$, where $\mathbf{1}$ is the all one vector with the same dimension of $\mathbf{M}_{D_a^*}$. Thus, we run a two-sample statistical test: we fix one sample to be $\mathbf{1}$ (an all-one vector), and use $\mathbf{M}$ as the second sample. We set the null hypothesis to be that *2 samples are drawn from the same distribution* (i.e., $\mathbf{M}$ is the sample-wise auditing results for a

Algorithm 2 Infer Membership Thresholds [15]

Input: A, the training algorithm; D_{cal}, the calibration dataset; $g_1, \cdots, g_m$, m different metrics for membership testing.
Output: $\tau_1, \cdots, \tau_m$, thresholds for m different metrics for membership inference.

procedure INFERTHRES
  Split D_{cal} into train and test datasets $D_{\text{cal}}^{\text{train}}$ and $D_{\text{cal}}^{\text{test}}$
  $f_{D_{\text{cal}}} \leftarrow A(D_{\text{cal}}^{\text{train}})$ ▷ Train the calibration model using the training subset of the calibration dataset
  for $i \in [m]$ **do**
    $\mathbf{V}_{\text{train}} \leftarrow \{g_i(f_{D_{\text{cal}}}, s) | s \in D_{\text{cal}}^{\text{train}}\}$ ▷ Compute metrics for the training subset of the calibration dataset
    $\mathbf{V}_{\text{test}} \leftarrow \{g_i(f_{D_{\text{cal}}}, s) | s \in D_{\text{cal}}^{\text{test}}\}$ ▷ Compute metrics for the testing subset of the calibration dataset
    $\tau_i \leftarrow \arg\max_{\tau \in [\mathbf{V}_{\text{train}}, \mathbf{V}_{\text{test}}]} (\text{BA}(\tau))$ ▷ Infer the threshold based on Eq 2
  end for
  return $\tau_1, \cdots, \tau_m$
end procedure

used audited dataset). The test will return a p-value, where a large p-value indicates weak evidence against (*i.e.,* supports) the null hypothesis that $\mathbf{M}$ is from the all-1 distribution. Thus, we use p-value as the final output of our EMA scheme, and we denote it as ρ_{EMA}. We interpret ρ_{EMA} as follow: if $\rho_{\text{EMA}} \leq \alpha$, the auditor can reject the null hypothesis, and conclude that the audited dataset is not used. Here, α is the threshold for statistical significance, and is set to 0.1 by default.

Differences from the previous method. Table I lists the differences between our method and CaliForget. As shown, our EMA addresses limitations of the previous method by avoiding possible false-negative (due to low quality calibration data) and false-positive cases (due to similar audited data to training data), which we are going to show in the next section. EMA is more cost-efficient since it does not require training a model on the audited dataset.

VI. EXPERIMENTS

We conduct experiments with two benchmark datasets, four chest X-ray datasets and two skin lesion datasets. We aim at evaluating how well EMA performs with respect to the requirements of dataset auditing for machine learning models I. Our experiments focus on the aspects of the practical and robust requirement and use previous work CaliForget as our baseline.

A. Experimental setups

We first evaluate EMA with benchmark datasets for proof-of-concept (Section VI-B). We then carry out comprehensive evaluations with medical image datasets, and study the robustness of EMA under various settings (Section VI-C and VI-D). Specifically, we want to evaluate the performance of EMA on various distributions of the audited dataset (C5), various audited dataset sizes (C6), and sensitivities to the qualities of the calibration dataset (C7).

We use the PyTorch framework [31] for deep learning implementations. All experiments use NVIDIA Tesla T4 GPUs and custom Intel Cascade Lake CPUs as provided by Amazon Web Services. We use a slightly different experimental setting for different tasks. We provide detailed experimental setup, model architectures and training configurations, for each dataset we evaluate in the following paragraphs.

Interpreting scores of EMA and CaliForget. EMA and CaliForget use different metrics at different scales to indicate whether the audited dataset is used in training the target machine learning model. We scale the auditing score of CaliForget to ease our comparisons of experimental results.

- EMA: auditing score ρ_{EMA} is in the range of $0 \sim 1$. If $\rho_{\text{EMA}} \leq \alpha$ (where $\alpha = 0.1$), the audited dataset is not used to train the targeted model.
- CaliForget: scale auditing score ρ_{CF} the range of $0 \sim 1$ and then define $\hat{\rho}_{\text{CF}} = 1 - \rho_{\text{CF}}$. If $\hat{\rho}_{\text{CF}} \leq 0.8$, the audited dataset is considered not used in training the targeted model.

By such scaling, we can compare the auditing scores of EMA and CaliForget easily, though they have different thresholds.

Evaluation setup. Our experiments evaluate with different image modalities, including handwritten digits (Section VI-B), chest X-ray images (Section VI-C), and skin lesion images (Section VI-D). For each image modality, we fix the training dataset $\mathcal{D}$, and vary the audited dataset D_{a} and the calibration dataset D_{cal}:

- We evaluate with three types of audited datasets: D_{a}^{tr}, a set of images used during training (*i.e.,* training images); D_{a}^{te}, a set of images that are drawn from the distribution of images in $\mathcal{D}$ but not used during training (*i.e.,* testing images); and $D_{\text{a}}^{\text{ood}}$, a set of images that are neither drawn from the distribution of images in $\mathcal{D}$ nor used during training (*i.e.,* out of distribution).
- We evaluate with two types of calibration datasets: $D_{\text{cal}}^{\text{id}}$, a set of images that are drawn from the distribution of images in $\mathcal{D}$ (*i.e.,* in distribution); and $D_{\text{cal}}^{\text{ood}}$ a set of images that are not drawn from the distribution of images in $\mathcal{D}$ (*i.e.,* out of distribution).

B. Results with Benchmark Datasets

We start by verifying the feasibility of EMA on benchmark datasets, MNIST dataset [32] which contains 60,000 images with image resolution 28×28, and SVHN dataset [33] which contains $73,257$ images in natural scenes with image resolution 32×32. The SVHN images are resized to 28×28 to be consistent with MNIST. We generate the training dataset, the calibration dataset, and the audited dataset as follows.

- **Training dataset** $\mathcal{D}$: we randomly sample $10,000$ images from MNIST as the training dataset and split them equally to 5 non-overlapping folds, and we have $\mathcal{D} = \{D_1, \cdots, D_5\}$.
- **Calibration dataset** D_{cal}: we sample $10,000$ MNIST images (disjoint with the training dataset) to construct the calibration dataset:
  1) $D_{\text{cal}}^{\text{id}}$: we use these $10,000$ original images as the in-distribution calibration dataset;

D_{cal}	k	D_{a}^{tr}	D_{a}^{te}	$D_{\text{a}}^{\text{ood}}$
$D_{\text{cal}}^{\text{id}}$	100	0.82	*0.84*	0.53
$D_{\text{cal}}^{\text{ood}}$	90	0.81	*0.83*	0.48
	80	0.81	*0.83*	0.25
	70	0.81	*0.82*	0.78
	60	**0.79**	*0.82*	0.05
	50	**0.79**	*0.81*	0.48

(a) ρ_{CF} scores of CaliForget.

D_{cal}	k	D_{a}^{tr}	D_{a}^{te}	$D_{\text{a}}^{\text{ood}}$
$D_{\text{cal}}^{\text{id}}$	100	1.00	0.00	0.00
$D_{\text{cal}}^{\text{ood}}$	90	1.00	0.00	0.00
	80	1.00	0.00	0.00
	70	1.00	0.00	0.00
	60	1.00	0.00	0.00
	50	1.00	0.00	0.00

(b) ρ_{EMA} scores of EMA

TABLE II: Auditing scores of both methods on **benchmark datasets**. Each column corresponds to an audited dataset, and each row corresponds to a calibration set with quality controlled by k. The false-negative results are in **red**, while the false-positive results are shown in *blue*.

2) $D_{\text{cal}}^{\text{ood}}$: To simulate the out-of-distribution calibration set in practice, we keep $k\%$ of the original images, add random Gaussian noise to $(100-k)/2\%$ of the images, and randomly rotate the other $(100-k)/2\%$ of the images. We vary k in our evaluation. Note that $k = 100$ implies $D_{\text{cal}}^{\text{id}}$.

- **Audited dataset** D_{a}: we test three audited datasets:
  1) D_{a}^{tr}: 5 folds of MNIST images used in training, each with $2,000$ images;
  2) D_{a}^{te}: $2,000$ images randomly selected from the MNIST dataset (disjoint with the training and the calibration set);
  3) $D_{\text{a}}^{\text{ood}}$: $2,000$ images randomly selected from the SVHN dataset.

Target model. The target model is a three-layer multi-layer perceptron of hidden size $(256, 256)$. Its training uses the SGD optimizer [34] with learning rate 0.05 (learning rate decay is set to 10^{-4}) and runs for 30 epochs with batch size 64.

Table II shows the auditing results of EMA and CaliForget. In our experiment, EMA obtained $\rho_{\text{EMA}} = 1$ for the MNIST folds used for training (*i.e.,* D_{a}^{tr}) and $\rho_{\text{EMA}} = 0$ for the MNIST fold not used for training, D_{a}^{te} under each clean data ratio $k\%$ for $k \in \{100, 90, 80, 70, 60, 50\}$, which gives correct auditing result. The baseline method CaliForget acquired correct ρ_{CF} values for D_{a}^{tr} when the clean data ratio $k \geq 70$ (*i.e.,* less noisy calibration data) and for $D_{\text{a}}^{\text{ood}}$ under every k. However, when k drops below 70, CaliForget get false-negative results (*i.e.,* $\rho_{\text{CF}} > 1$) for D_{a}^{tr}. As for D_{a}^{te}, CaliForget returns false-positive results since the audited dataset is similar to the training dataset.

C. Results with X-ray datasets

As we demonstrate in Sec. VI-B, EMA achieves consistently superior results in different settings, we then evaluate EMA on medical imaging tasks on Chest X-ray datasets for pneumonia patient v.s. healthy control binary classification . We are especially curious about:

1) The **performance** of EMA on medical data such as Chest X-ray;
2) The **robustness** of EMA with various numbers of **audited datapoints**;

X-Ray Dataset		**#Datapoints**		**Accuracy**	
		Train	Test	Train	Test
$\mathcal{D}$	COVIDx [35]	4,000	1,000	1.00	0.89
$D_{\text{cal}}^{\text{id}}$	COVIDx, k=100	4,000	1,000	1.00	0.90
$D_{\text{cal}}^{\text{ood}}$	COVIDx, k=80	4,000	1,000	0.98	0.88
	COVIDx, k=60	4,000	1,000	0.98	0.86
	CXR-NIH [36]	15,000	5,000	1.00	0.70
	CBIS-DDSM [37]	2,676	336	1.00	0.84

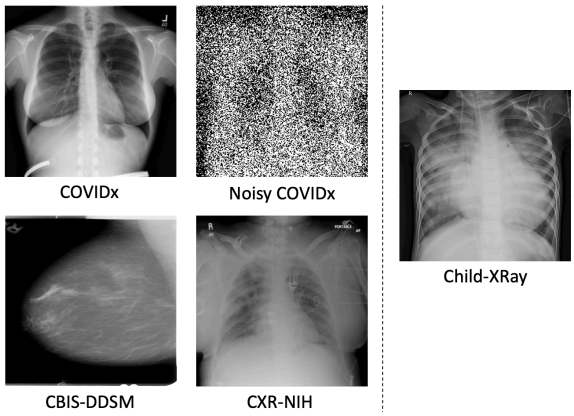

TABLE III: Table: detailed information of the X-ray datasets. Figures: example images of our used chest X-ray datasets. Left hand side are our D_a^{ood}'s, and right hand side is our D_a^{ood}.

3) The **capacity** of EMA using **calibration datapoints** with different quality.

Experimental Setup. We first validate EMA's performance (Sec. VI-C.1) with X-ray datasets, and then evaluate EMA's robustness (Sec. VI-C.2) and capacity (Sec. VI-C.3) by varying its configurations for audited and calibration datasets.

Similar to the setup in Sec. VI-B, we split the X-ray datasets into training, calibration, and audited datasets.

- **Training dataset**: We randomly sample $4,000$ images from COVIDx [38] to train the target model. Since the other chest X-ray datasets we used are for binary classification, we merge COVID19 and pneumonia into one class in COVIDx and perform binary classification to align with binary labels on the other X-ray datasets.
- **Calibration dataset** D_{cal}:
  1) $D_{\text{cal}}^{\text{id}}$: we sample another $4,000$ images from

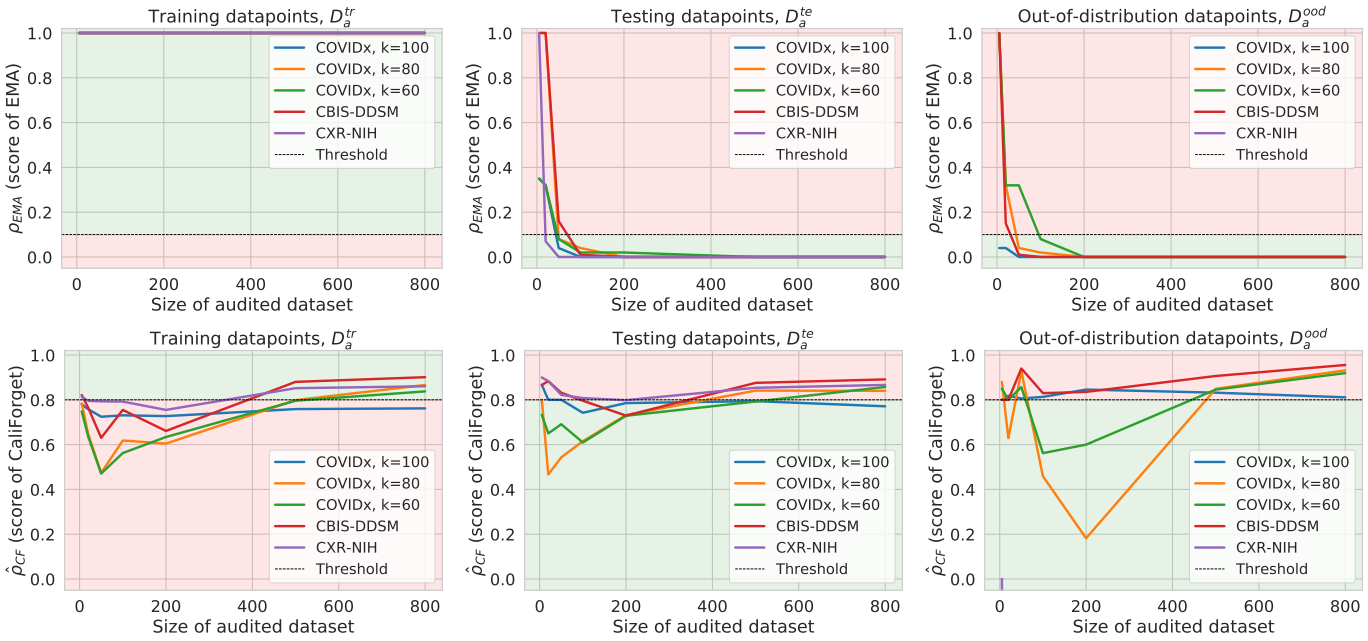

Fig. 3: ρ_{EMA} (top) and $\hat{\rho}_{\text{CF}}$ (bottom) scores for Chest X-ray experiments. Green regions mark the expected (or correct) auditing scores while red regions mark the wrong auditing scores: for D_{a}^{tr}, we expect the auditing score to be higher than the threshold (*i.e.,* predict the dataset as 'having been used during training'), while for D_{a}^{te} and $D_{\text{a}}^{\text{ood}}$ we expect otherwise. EMA gives correct auditing scores for D_{a}^{tr} regardless of the size of audited dataset, while for D_{a}^{te} and $D_{\text{a}}^{\text{ood}}$, the performance of EMA improves as the the size of audited dataset increases (EMA achieves correct dataset auditing for $|D_a| \geq 100$). On the other hand, for CaliForget, there are several false-negative $\hat{\rho}_{\text{CF}}$ scores for D_{a}^{tr} and false-positive $\hat{\rho}_{\text{CF}}$ scores for D_{a}^{te} and $D_{\text{a}}^{\text{ood}}$.

COVIDx (disjoint with the training dataset) as the in-distribution calibration dataset;

2) $D_{\text{cal}}^{\text{ood}}$: we also evaluate out-of-distribution calibration datasets, including the COVIDx dataset with noise[4], and other X-ray datasets (see Table III and Section VI-C.3 for details).

- **Audited dataset** D_{a}: we select the following public available X-ray datasets and analog them as three types of audited dataset:
  1) D_{a}^{tr}: 5 non-overlapping folds of COVIDx images used in training, each with 800 images;
  2) D_{a}^{te}: $1,000$ images randomly selected from the COVIDx dataset (disjoint with the training and the calibration set);
  3) $D_{\text{a}}^{\text{ood}}$: 800 images randomly selected from the Child-XRay dataset [39].

We use ResNet50 [40] as the classification model and train it with Adam optimizer using ExponantialLR scheduler. We set the initial learning rate to 0.001, and train for 50 epochs with batch size 64. Early-stopping is applied to avoid severe overfitting. The data pre-processing in this section is fixed, for which we resize the images into 224×224, and normalize them to $\{mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]\}$ for each channel. The detailed information about the datasets and training results can be found in Table.III.

Experimental Results. For better illustration, we visualize EMA and CaliForget scores as shown in Fig.3. Since we have three types of audited datasets, we generate individual heatmap for each of them. For the X-ray data used for training (*i.e.,* D_{a}^{tr}), we report their averaged auditing scores; as for the ones not used for training (*i.e.,* D_{a}^{te} and $D_{\text{a}}^{\text{ood}}$), we report their original EMA scores. Recall that when $\rho_{\text{EMA}} \leq \alpha$ ($\alpha = 0.1$) and $\rho_{\text{CF}} \leq 0.8$, the audited data is considered not used by the model. Therefore, for ρ_{EMA}, we mark blue for the values ≤ 0.1 and red otherwise, whereas similarly, for $\hat{\rho}_{\text{CF}}$, we mark blue for the values ≤ 0.8 and red otherwise. In general, a more red-ish entry corresponds to a more confident prediction of 'the audited dataset is used', and a more blue-ish entry corresponds to a more confident prediction of 'the audited dataset is not used'. For the following experiments, we use t-test for our statistical test.

1) Performance of EMA *with X-ray images:* We present the performance of EMA and CaliForget with X-ray images in Fig.3. For the default setup where COVIDx(k100), *i.e.,* the original COVIDx dataset without noisy data as calibration dataset and audited size = 800, ρ_{EMA} are all 1's for D_{a}^{tr} and 0's for D_{a}^{te} and $D_{\text{a}}^{\text{ood}}$, which indicates that EMA is able to distinguish not only the data from different source (*i.e.,* $D_{\text{a}}^{\text{ood}}$) but also data from the same dataset but not included in base model (*i.e.,* D_{a}^{te}). For the evaluation on CaliForget, we expect to see $\hat{\rho}_{\text{CF}} \leq 0.8$ for datasets that have been used in training and otherwise for unused datasets, under the default setting (COVIDx(k100) and audited size = 800). However, many of the $\hat{\rho}_{\text{CF}}$ values in D_{a}^{tr} are smaller than 0.8, which gives us

[4]We keep $k\%$ of the $4,000$ $D_{\text{cal}}^{\text{id}}$ images, and add random Gaussian noise to $(100-k)\%$ of the images.

false-negative results. Besides, for $\hat{\rho}_{\mathrm{CF}}$ in $D_{\mathrm{a}}^{\mathrm{ood}}$, it is larger than 0.8, which gives a *false-positive* result. Consequently, CaliForget is relatively unstable for auditing COVIDx models.

2) Robustness of EMA under different audited sizes: To examine the robustness of EMA with different number of audited datapoints ($|D_a|$), we fix the clean data ratio $k = 100$, and vary the audited dataset size $|D_a|$ from 800 to 5. The first rows of each ρ scores in Fig.3 show the auditing scores for each audited dataset size. It is noted that ρ_{EMA}'s for $D_{\mathrm{a}}^{\mathrm{tr}}$ is always one (> 0.1), which means they are correctly considered used in training the model. For $|D_a| = 20$, ρ_{EMA} for $D_{\mathrm{a}}^{\mathrm{te}}$ is 0.02, and ρ_{EMA} for $D_{\mathrm{a}}^{\mathrm{ood}}$ is 0, which means they are correctly considered unused. Based on the result, we conclude that EMA is robust when $|D_a| > 20$. However, CaliForget returns several false-negative(blue entries in $D_{\mathrm{a}}^{\mathrm{tr}}$) and false-positive (red entries in $D_{\mathrm{a}}^{\mathrm{te}}$ and $D_{\mathrm{a}}^{\mathrm{ood}}$) results regardless of the chosen $|D_a|$.

3) Capacity of EMA under various calibration datasets: We further investigate EMA's capacity under different calibration datasets. Here, we test 4 additional calibration datasets:

- **Noisy COVIDx**: Similar to Sec.VI-B, we add noise controlled by k to the dataset. However, for this time we use only $k = 20, 40$ for simplicity.
- **CXR-NIH** [36]: A chest X-ray dataset with 8 different type of diseases. The motivation for this is to test EMA with a different-purposed X-ray dataset as calibration dataset. The 2 classes extracted are *Atelectasis* and *No Finding*.
- **CBIS-DDSM** [37]: A breast mammogram X-ray dataset for breast cancer classification. The motivation is to test EMA with mammography X-ray dataset as calibration dataset.
- **The audited dataset itself**: the motivation of this is that we assume the audited dataset is all we have for auditing. We want to test if EMA is still applicable in this case.

We also vary the audited dataset size $|D_a|$ in the experiments. Table III provides the training and testing accuracy of different calibration models. As shown in Fig.3, EMA works well for all the cases in $D_{\mathrm{a}}^{\mathrm{tr}}$, and only fails in $D_{\mathrm{a}}^{\mathrm{te}}$ and $D_{\mathrm{a}}^{\mathrm{ood}}$ when $|D_a| \leq 50$. More explicitly, $D_{\mathrm{a}}^{\mathrm{te}}$ is our bottleneck test case, which is the most difficult audited dataset for dataset auditing since it's very similar to the training dataset. Consequently, the results show that EMA can achieve correct dataset auditing for $|D_a| > 50$ under the COVIDx experiment.

D. Results with Dermatology datasets

To further extend EMA to color medical datasets, we experiment EMA on dermatology images. We construct training, calibration, and audited datasets based on HAM10000 [41], a multi-source dermatology dataset for skin lesion classification.

- **Training dataset** : we randomly sample 600 images from HAM10000 (HAM in the following for simplicity) as the training dataset.
- **Calibration dataset** D_{cal}:
  1) $D_{\mathrm{cal}}^{\mathrm{id}}$: we randomly sample another 600 images from HAM10000 (disjoint with the training dataset) as the in-distribution calibration dataset;
  2) $D_{\mathrm{cal}}^{\mathrm{ood}}$: we also evaluate out-of-distribution calibration datasets. One of them is a publicly available skin lesion dataset (*i.e.,* PAD-UFES-20 [42]). To further test if natural RGB images can serve as calibration datasets (which are easier to acquire in practice), we also evaluate with a natural image dataset CIFAR-10 [43].
- **Audited dataset** D_{a}: we design the following three kinds of audited dataset:
  1) $D_{\mathrm{a}}^{\mathrm{tr}}$: 3 non-overlapping folds of HAM10000 images used in training, each with 200 images;
  2) $D_{\mathrm{a}}^{\mathrm{te}}$: 200 images randomly selected from the HAM10000 dataset (disjoint with the training and the calibration set);
  3) $D_{\mathrm{a}}^{\mathrm{ood}}$: 137 images randomly selected from the DERM7pt [44] dataset.
- **PAD-UFES-20** [42]: A skin lesion dataset collected from smartphones for 6 diagnostics classification. The motivation is to test whether using dataset with similar purpose but taken with different imaging methodology can still work as calibration dataset for EMA auditing.
- **CIFAR-10** [43]: A RGB color image dataset for natural images classification. The motivation of using CIFAR-10 as the calibration dataset is to test if model trained with *out-of-domain* datasets can still infer accurate auditing score.

Since the pathology from the datasets are unevenly distributed and not fully-overlapped, we extract the common 3 classes (*seborrheic keratosis*, *basal cell carcinoma* and *nevus*) from HAM10000 and PAD-UFES-20, and train 3-class classifiers for them. Accordingly, we randomly select three classes (dogs, cats and birds) from CIFAR-10 to match the 3-class classifi-

Dermatology Dataset		#Datapoints		Accuracy	
		Train	Test	Train	Test
$\mathcal{D}$	HAM10000 [41]	600	100	1.00	0.90
$D_{\mathrm{cal}}^{\mathrm{id}}$	HAM10000	600	100	0.99	0.92
$D_{\mathrm{cal}}^{\mathrm{ood}}$	PAD-UFES-20 [42]	600	120	0.99	0.67
	CIFAR-10 [43]	15,000	3,000	0.99	0.78

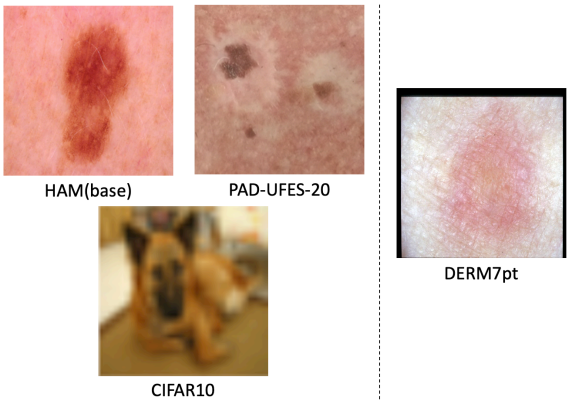

TABLE IV: Detailed information of the dermatology datasets and some example images. Left hand side are our D_a^{ood}'s, and right hand side is our D_a^{ood}.

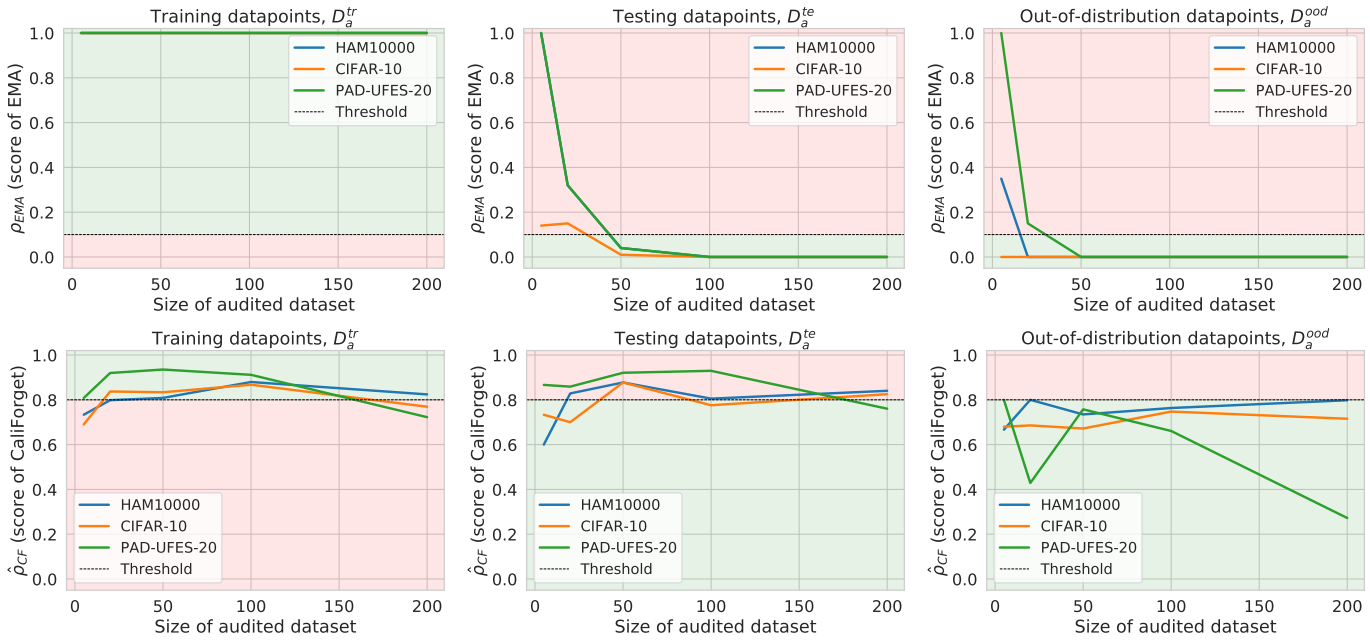

Fig. 4: ρ_{EMA} (top) and $\hat{\rho}_{\text{CF}}$ (bottom) scores for Dermatology experiments. Green regions mark the expected (or correct) auditing scores while red regions mark the wrong auditing scores: for D_{a}^{tr}, we expect the auditing score to be higher than the threshold (*i.e.,* predict the dataset as 'having been used during training'), while for D_{a}^{te} and $D_{\text{a}}^{\text{ood}}$ we expect otherwise. EMA gives correct auditing scores for D_{a}^{tr} regardless of the size of audited dataset, while for D_{a}^{te} and $D_{\text{a}}^{\text{ood}}$, the performance of EMA improves as the the size of audited dataset increases (EMA achieves correct dataset auditing for $|D_a| \geq 50$). On the other hand, although CaliForget correctly finds out $D_{\text{a}}^{\text{ood}}$ has not been used by the target model, there are several false-negative $\hat{\rho}_{\text{CF}}$ scores for D_{a}^{tr} and false-positive $\hat{\rho}_{\text{CF}}$ scores for D_{a}^{te}.

cation task of skin lesion analysis task.

Results of EMA (*i.e.,* ρ_{EMA}) and CaliForget (*i.e.,* $\hat{\rho}_{\text{CF}}$) scores are shown in Fig.4. Note that in the top row, EMA can successfully audit the usage of all audited datasets using any of the calibration datasets when $|D_a| = 50$. For the bottom row, there are several false-negative ρ_{CF} scores in D_{a}^{tr} and false-positive $\hat{\rho}_{\text{CF}}$ scores in D_{a}^{te} for CaliForget, but it gives correct ρ_{CF} scores in $D_{\text{a}}^{\text{ood}}$ under all $|D_a|$'s. The results show that EMA can achieve correct dataset auditing for $|D_a| \geq 50$ under the dermatology experiment.

VII. Conclusion

This paper first presents a set of requirements for dataset auditing for collaboratively training ML models with multiple datasets in a Federated Learning setting. Then, we describe EMA, a dataset auditing method based on membership inference attack and statistical ensemble.

In terms of time efficiency, EMA is superior to the previous state-of-the-art CaliForget, as it does not require training a model with the audited dataset. When the audited dataset is large, the time saving could be significant.

Our extensive experiments on benchmark and medical datasets show that EMA works well with all datasets. It works robustly with as small as 50 data points in the audited dataset size, whereas CaliForget fails when there are fewer than 800 data points for certain medical image datasets.

Similar to CaliForget, EMA requires finding a publicly available dataset as its calibration dataset, although EMA is much more robust with different qualities of calibration datasets than CaliForget. A future direction is to either automatically generate a calibration dataset using synthetic dataset generate algorithms, or develop suitable metrics to decide whether a given dataset is a good candidate for a calibration dataset.

We also would like to derive a theoretical guarantee for EMA, which is important for its real-world deployment. We appreciate the interesting theoretical analysis in a recent work [45]. However their results are limited to linear classifiers, which makes a direct comparison with ours difficult, as our empirical studies have been carried out with real-world convolutional neural networks. Ultimately, we expect this work to improve the awareness of data privacy, increase trustworthiness, and accelerate the development of applications in AI for healthcare.

References

[1] A. Act, "Health insurance portability and accountability act of 1996," *Public law*, vol. 104, p. 191, 1996.
[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, 2017.
[3] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," 2019. [Online]. Available: https://arxiv.org/abs/1912.04977
[4] C. S. Legislature, "California consumer privacy act," 2018. [Online]. Available: https://oag.ca.gov/privacy/ccpa
[5] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," *Intersoft consulting*, 2018.

[6] X. Liu and S. A. Tsaftaris, "Have you forgotten? a method to assess if machine learning models have forgotten data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 95–105.

[7] S. Barkow and K. Takahashi, "Current expectations and guidance, including data integrity and compliance with cgmp," *Center for Drug Evaluation and Research*, 2017.

[8] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[9] Y. Huang, X. Li, and K. Li, "Ema: Auditing data removal from trained models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 793–803.

[10] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, 2017, pp. 587–601.

[11] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 267–284.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[13] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[14] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.

[15] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," *arXiv preprint arXiv:2003.10595*, 2020.

[16] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*, 2015, pp. 463–480.

[17] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou, "Making ai forget you: Data deletion in machine learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[18] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 3832–3842. [Online]. Available: https://proceedings.mlr.press/v119/guo20c.html

[19] S. Garg, S. Goldwasser, and P. N. Vasudevan, "Formalizing data deletion in the context of the right to be forgotten," in *Advances in Cryptology – EUROCRYPT 2020*, A. Canteaut and Y. Ishai, Eds. Cham: Springer International Publishing, 2020, pp. 373–402.

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

[21] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 896–911. [Online]. Available: https://doi.org/10.1145/3460120.3484756

[22] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites, "Adaptive machine unlearning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 16 319–16 330. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/87f7ee4fdb57bdfd52179947211b7ebb-Paper.pdf

[23] A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh, "Remember what you want to forget: Algorithms for machine unlearning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 18 075–18 086. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/9627c45df543c816a3ddf2d8ea686a99-Paper.pdf

[24] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *ACM Conference on Computer and Communications Security (CCS)*, 2020.

[25] H. A. Inan, O. Ramadan, L. Wutschitz, D. Jones, V. Rühle, J. Withers, and R. Sim, "Training data leakage analysis in language models," *arXiv preprint arXiv:2101.05405*, 2021.

[26] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[27] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[28] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[29] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 634–646.

[30] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1605–1622.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[32] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[33] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[34] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[35] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[36] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[37] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[38] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, p. 19549, Nov 2020.

[39] D. S. Kermany, K. Zhang, and M. H. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," 2018.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[42] A. G. Pacheco, G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro *et al.*, "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in brief*, vol. 32, p. 106221, 2020.

[43] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, vol. 55, no. 5, 2014.

[44] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019.

[45] C. Yadav, M. Moshkovitz, and K. Chaudhuri, "A learning-theoretic framework for certified auditing of machine learning models," *arXiv preprint arXiv:2206.04740*, 2022.