Certified in Theory, Broken in Practice: Assumption Gaps in Cryptographic Model Certification

Carter Luck[*]
University of Massachusetts Amherst

Olive Franzese-McLaughlin[*]
Vector Institute & University of Toronto

Elisaweta Masserova[*]
Carnegie Mellon University

Akira Takahashi
J.P.Morgan AI Research & AlgoCRYPT CoE

Antigoni Polychroniadou
J.P.Morgan AI Research & AlgoCRYPT CoE

Nicolas Papernot
Vector Institute & University of Toronto

Abstract

Privacy-preserving machine learning auditing protocols allow auditors to assess models for properties such as accuracy or fairness, without revealing their internals or training data. This makes them especially attractive for auditing models deployed in sensitive domains such as healthcare or finance.For these protocols to be meaningful in real-world audit settings, though, their guarantees must reflect how the model will behave once deployed, rather than merely certifying its behavior during an audit. Existing security definitions often miss this mark: most certify model behavior only on a *fixed audit dataset*, without ensuring that the same guarantees *generalize* to other datasets drawn from the same distribution. As we show, this gap allows a model provider to attack many cryptographic model certification (CMC) schemes built on secure zero knowledge proofs (ZKP) by carefully engineering training data, resulting in models that exhibit benign behavior during an audit, but pathological behavior in practice. For example, we empirically demonstrate that an attacker can certify that a model achieves over 99% accuracy on an audit dataset, but less than 30% accuracy on fresh samples from the same distribution.

To address this gap, we formalize rigorous cryptographic security notions tailored to CMC frameworks, introduce a generic protocol template, and prove that it satisfies these requirements. Our results thus offer both cautionary evidence about existing approaches and constructive guidance for designing secure, privacy-preserving ML auditing protocols.

1 Introduction

Certifiable, privacy-preserving machine learning aims to formally prove desired properties of the model while keeping model parameters and training data confidential [35, 42, 56]. In this context, the typical certification lifecycle follows a sequence in which the model provider first trains the model, then an auditor evaluates it according to desired criteria, and—after passing the audit—the certified model is deployed.[1] Crucially, the trustworthiness of this certification does not rely on trusting the model provider or a specific machine learning algorithm. Instead, it is ensured by cryptographic mechanisms – such as commitments and zero-knowledge proofs [26] – which guarantee that the auditor's evaluation is performed correctly on the claimed model. The usage of cryptographic techniques allows the auditor to not only verify whether the model has the intended property, but do so while keeping the model internals and training data private.

However, as we observe, the guarantees that *cryptographic model certifications* (CMC) provide often fail to reflect the model's behavior beyond the specific setting in which certification is performed. In this paper, we show that such audit-specific guarantees risk creating a false sense of security: by themselves, they do not ensure that the certified properties will continue to hold once the model is deployed and applied to *fresh data*, even when this data is drawn from the same distribution as the audit dataset. We demonstrate that this is not merely a theoretical concern.

In more detail, consider a certification workflow, where the audit dataset is drawn from public benchmarks, or derived from distributions that providers can closely approximate. In such scenarios, when the audit dataset is known to the model provider – even partially – prior to the start of the audit procedure[2], we describe simple yet powerful attack strategies that can be executed by an adversarial model provider. Our attacks allow the provider to pass an audit (thus enabling deployment) while simultaneously pursuing its own, potentially conflicting, interests. For example, while an auditor may seek to verify fairness, the model provider may instead prioritize accuracy—even when accuracy and fairness are in tension. As a result of the attack, the audited model passes an audit without exhibiting the intended behavior in practice, and this holds

[*]Equal contribution

[1]Some works require continuous auditing during deployment instead of a single audit pre-deployment; see Table 1.

[2]Or, alternatively, if the audit procedure does not require an audit dataset at all, such as in [41].

Table 1: Analysis of vulnerabilities to data-forging attacks in privacy-preserving ML audits. ✓= supported; ▲= conditional; ✗= not supported.

Work	Certified property				Resilience to data-forging	Continuous verification
	Acc.	Group Fair	Indv. Fair	Diff. Priv.		
Zhang et al. [56]	✓	✗	✗	✗	▲ (pd)	✗
Shamsabadi et al. [42]	✗	✓	✗	✗	✗	✗
Yadav et al. [51]	✗	✗	✓	✗	✓	✓
Liu et al. [35]	✓	✗	✗	✗	▲ (pd)	✗
Franzese et al. [21]	✗	✓	✗	✗	✓	✓
Shamsabadi et al. [41]	✗	✗	✗	✓	✗	✗
Kang et al. [33]	✓	✗	✗	✗	✓	✗
Wang and Hoang [48]	✓	✗	✗	✗	▲ (pd)	✗
Bourrée et al. [9]	✗	✓	✗	✗	✗	✗

Acc. = accuracy; Group/Indv. Fair = group/individual fairness; Diff. Priv.=differential privacy. "Conditional" works lack detail to assess resilience to data-forging, but indicate deployments with public datasets (pd), which would be make the solution vulnerable. Continuous verification means audits must run continuously during deployment (e.g., via clients) rather than once pre-deployment.

even if strong cryptographic tools are used to ensure that the model during the audit procedure is *the same* as the one used during the deployment. Our threat model reflects common certification practice (both in the academic literature and in practice). As surveyed in Appendix E, many prior works in model certification utilize publicly known datasets for audit purposes. Moreover, when a company audits a model provided by another company, the audit dataset—whether sourced from the company's synthetic test portfolio or from public benchmarks—is typically disclosed to the model provider, with no mechanism in place to ensure that providers finalize their trained models *before* seeing the audit data.

The key idea in our attacks that allows us to bypass common cryptographic defenses is that we do not alter the training process itself. Instead, we show that knowing the certification procedure enables the model provider to forge training data so that a model honestly trained on it passes the audit, yet still exhibits pathological behavior on other datasets coming from the same distribution as the audit dataset. We demonstrate the effectiveness of our attacks against certification procedures for accuracy, fairness, and differential privacy.

We empirically show that our **data forging attacks** can cause dramatic gaps between audit-time guarantees and true model performance: for instance, in one of our attacks a model can pass an audit requiring 99% accuracy on the audit dataset, yet achieve only 30% accuracy on new samples from the same distribution. We show that our attacks remain undetected by straightforward approaches such as statistical tests, e.g., Welch's t-test [49] are performed to check whether the training data and audit data were taken from the same distribution.

To assess how widespread vulnerability to data forging attacks may be, we conduct a case study covering nine recent papers published at top venues in both security and machine learning. We find that six of these works are at least partially vulnerable[3]. We provide a summary of results in Table 1.

Given these vulnerabilities, we introduce a formal foundation for cryptographic model certification that provides rigorous guidance for secure auditing practice. This foundation includes a security definition ensuring that a model provider passes the audit if and only if the model satisfies the desired property (up to a small quantifiable error) on the underlying data distribution. We further formalize an attack game that highlights the gap between certifying a property on a fixed dataset and certifying that the same property generalizes to fresh samples from the distribution. Finally, we propose a generic template for achieving secure CMC. Our approach is agnostic to the specific property that is being certified and, as we formally prove, guarantees that whenever a model passes an audit, the certified properties will also hold at deployment. The key ingredient is ensuring that the audit is conducted on test data that is *independent* of both the model and its training data. This method might serve as a template for future works to obtain not only efficient, but also secure auditing solutions. Applying this method allows to restore/expand security in some of the case study works.

In summary, our work advances the study of cryptographic auditing for machine learning by **(i)** identifying simple, yet effective, attack strategies that allow to pass an audit while enabling pathological model behavior at deployment with respect to new datasets from the same distribution as the audit one, **(ii)** empirically demonstrating the effectiveness of our attack against generic certification approaches for four objectives: accuracy, fairness, differentially private training, and statistics for distribution similarity testing, **(iii)** conducting a

[3]Several works do not specify the auditing protocol or formal security guarantee in sufficient detail to allow a definitive assessment. However, many explicitly consider deployments based on public audit datasets, under which their approaches are vulnerable.

systematic case study of prior certification approaches and corresponding threat models proposed in the literature and identifying several existing approaches that are susceptible to data forging vulnerabilities, and **(iv)** introducing formal security definitions tailored to certifiable machine learning and a protocol template that mitigates the attack. We emphasize that we *do not suggest that prior cryptographic works are broken on a technical level*, rather that the guarantees these works provide deserve closer scrutiny. Our findings comprise strong evidence that secure audit solutions with any of the following properties are unlikely: a) those which utilize known public datasets for test purposes, and b) those that reuse test datasets (if model owner learns a substantial amount of this test dataset during the audit). This evidences the importance of continuous sampling of fresh data for a successful audit infrastructure. We hope that our work will inform the design of future cryptographically secure machine learning audit frameworks.

2 Threat Model: Cryptographic Model Certification (CMC)

We consider two parties: a malicious model provider a.k.a. prover, and an honest-but-curious auditor[4] a.k.a. verifier. The auditor and model provider are in agreement on an interactive audit protocol wherein both input data (some of which may be *proprietary* data private to the model provider) and the model provider additionally inputs a set of model parameters. The auditor's goal is to assess whether the model has trustworthy properties such as accuracy, fairness, or differential privacy, while maintaining the confidentiality of the model and proprietary data. The protocol will output a bit indicating whether the model provider "passes" the audit.

The model provider may submit arbitrary model parameters, and arbitrary data to the protocols requested by the auditor. In some scenarios in the paper, we consider a model provider who can gain access to some or all of the auditor's data (mimicking many real-world scenarios where audit data is reused) prior to the audit. The goal of the model provider is to pass the audit, while exploiting their hidden information to surreptitiously embed the model with nefarious properties.

Cryptographic Model Certification vs. Data Poisoning. Attacking cryptographic model certification is related, but distinct from, the problem of data poisoning [44]. Such attacks have traditionally been considered in the context of ML systems trained on user-provided data. Both data poisoning and our attacks on CMC (see §4.1) involve adversarial manipulations of training data. However, their **threat models differ**, leading to fundamentally different attacks and defenses:

In data poisoning the model provider is the *defender*, and it seeks to learn an accurate model. The *attacker* is an external party who aims to corrupt the learned model by inducing the provider to incorporate an auxiliary training dataset D_p of size no larger than clean dataset D_c [6]. In contrast, in CMC the roles are reversed: **the attacker is the AI/ML service provider itself**, while the defender is an external auditor party. The goal of the service provider is to engineer a model that passes an audit, while violating the certified properties on real-world data. This scenario considers a vastly different set of adversarial capabilities, and imposes additional constraints on the defender:

1. A CMC adversary can arbitrarily manipulate both the model and the training data. This includes both label perturbation and feature-level manipulations. It further encompasses adversaries corrupting all training data, i.e., it *allows for traditional poisoning attacks, but without any restrictions on the poisoning ratio*. In fact, an adversarial model provider might simply chose not to include any clean data at all. Finally, the adversary may even corrupt the model parameters directly. Note that simply asking a model provider to execute a certain protocol (including a data poisoning defense mechanism) is not sufficient – a malicious model provider could choose to simply ignore this instruction.

2. Both the model and the training data are hidden from the defender. In contrast to poisoning threat models where the defender has full access to the model and data, the defending party in CMC is an external auditor, who is not allowed to view the model and data as they are typically proprietary. [5] This imposes additional constraints on the defender, who is tasked with assessing the validity of the model while keeping its parameters confidential. This is why known solutions with provably secure guarantees (including ours) crucially rely on cryptographic techniques.

To summarize, a CMC adversary can make use of classical poisoning techniques, including large-scale perturbations, label corruption, and visible backdoor triggers, to mount its attacks. However, the CMC adversary is strictly stronger than the adversary typically considered in traditional poisoning: It can use *additional* attack strategies that involve, e.g., arbitrarily modifying all data and even directly setting model parameters. Furthermore, the CMC's goal is fundamentally different: it does not attempt to detect or mitigate poisoned data, but instead ensure that any model that passes the audit

[4] We focus on honest-but-curious auditors for simplicity: it allows us to prove zero-knowledge (Theorem 3) from the weakest possible assumption, namely an honest verifier ZK. Our main focus is on the soundness gap caused by malicious model providers, so this choice does not affect the core contributions. This is not an inherent limitation: if the underlying proof system is ZK against fully malicious verifiers (e.g., a non-interactive proof system in the common reference string or random oracle model), then Π_{csp} achieves ZK against malicious auditors by the same argument.

[5] CMC can also be applied to auditing openly accessible (e.g., open-source) models, where confidentiality of model weights is not a concern. In this setting the system can be made more efficient: one can use a succinct proof system without the zero-knowledge property to reduce verification overhead, and skip generating cryptographic commitments to model weights and the corresponding commitment-opening proofs. The soundness guarantees of this work apply in this setting as well.

must satisfy the audited property, regardless of how it was produced.

Cryptographic Model Certification and Public Benchmarks in ML. The conclusions we draw about requiring fresh data for auditing are semantically related to work on the inadequacy of public benchmarks in machine learning [29,55], but those works do not consider cryptographic security. For additional related work and an overview of certifiable ML, see §A.

In §5, our goal is to design a CMC such that – even under these powerful adversarial capabilities – if the model passes the audit, then it must genuinely satisfy the audited property (e.g., a fairness or accuracy threshold).

3 Background

Consider the following scenario: An auditor wishes to verify whether a model utilized by an insurance company to justify claim decisions (approve/deny claim) is accurate on a dataset of the auditor's choosing. At the same time, the company does not want to reveal its model due to concerns about privacy and business competition. In CMC cryptographic techniques are deployed to reconcile these seemingly conflicting goals.

3.1 Zero Knowledge Proofs

Among these techniques, the central tool is *zero-knowledge (ZK) proofs* [26], a classical cryptographic primitive, which allows one party (a *prover*) prove a statement x to another party (*verifier*) without revealing anything else apart from the validity of this statement. Such proofs are constructed for a concrete NP relation $\mathcal{R}$, which is used to formalize what it means for a statement to be true by specifying the type of evidence (witness w) that certifies it. The statement x is public, the witness w is private, and the zk proof checks $(\mathsf{x},\mathsf{w}) \in \mathcal{R}$, without revealing w. In certifiable ML, such proofs allow model provider (prover) to formally prove that a model (witness) satisfies a desired property (e.g., accuracy, fairness, or inference correctness) on a given test dataset (statement) without learning anything else about the model or the training data. More formally:

Definition 1 (Proof System)**.** *An (interactive)* proof system ZKP *for an NP relation* $\mathcal{R}$ *is a tuple of interactive Turing machines* $(\mathcal{P},\mathcal{V})$*, where* $\mathcal{P}$ *is prover and* $\mathcal{V}$ *is verifier. Let* $b \leftarrow \langle \mathcal{P}(\mathsf{w}), \mathcal{V}\rangle(\mathsf{x})$ *denote the interaction between* $\mathcal{P}$ *and* $\mathcal{V}$*, where both* $\mathcal{P}$ *and* $\mathcal{V}$ *take* x *as common inputs, and* $\mathcal{P}$ *additionally takes* w *as a private input. At the end of interaction,* $\mathcal{V}$ *halts by outputting a binary* b*.*

Proof systems that are used in ML auditing typically require the following security properties: For an NP relation $\mathcal{R}$, they must provide *completeness* (i.e., if the prover and verifier follow the protocol with input $(\mathsf{x},\mathsf{w}) \in \mathcal{R}$, the verifier always accepts), *knowledge soundness* (i.e., if the verifier accepts, then it must be that the prover owns a valid witness w satisfying the given NP relation w.r.t. statement x), and *zero knowledge* (i.e., the transcript of the interaction between an honest prover and an honest verifier leaks nothing except that there exists a witness w such that $(\mathsf{x},\mathsf{w}) \in \mathcal{R}$). We defer formal definitions to §B.3.

3.2 Cryptographic Commitments

Returning to our example, suppose the insurance company has successfully passed an audit and can now deploy its model. How can a customer submitting inference queries be assured that the company continues to use the *certified* model—rather than switching to a different, unverified one? Again, the company still wishes to keep its model private. The standard cryptographic tool here is *commitments*, which bind the provider to a single private model during the audit. This prevents "model switching" and ensures that model used in deployment is the same as the one that was certified:

Definition 2 (Commitment Scheme)**.** *A commitment scheme is an algorithm* Commit*, which is executed as* $\mathsf{com} \leftarrow \mathsf{Commit}(m;\rho)$*. It takes as input a message* m*, a uniformly sampled randomness* ρ*, and returns a commitment* com*.*

We require two security properties: *hiding* (i.e., given a commitment com, it leaks nothing about the message m), and *binding* (i.e., it is computationally infeasible to open the same com to distinct m and m'). We defer formal definitions to §B.4. Now, auditing may require publishing such a commitment to the model,[6] after which the client and insurance company engage in a ZK proof of inference against it[7]. Figure 1 shows the full certification workflow, where the audit dataset is revealed to the model provider prior to committing to the model.

3.3 Unifying Syntax for Prior CMC Schemes

We will next discuss the security guarantees of works that address the certification stage (Figure 1, Left)—namely, proofs of accuracy, fairness, etc., between auditor and model provider. To analyze these systematically, rather than case by case, we abstract away implementation details and introduce a unifying syntax that captures a broad class of existing auditing systems [9,35,41,42,48,56]. To this end, we introduce a *certification predicate* f outputting binary depending whether the input satisfies certain properties. Then given a commitment scheme Commit and f, we can define a ZKP for the following

[6] The commitment may be signed by the auditor.

[7] In more detail, the insurance company will supply the inference outcome y along with a ZK proof proving that 1) the inference outcome is y has been computed using a private model M and the given client's input, and 2) M is the model hidden behind the previously published commitment that was approved by the auditor.

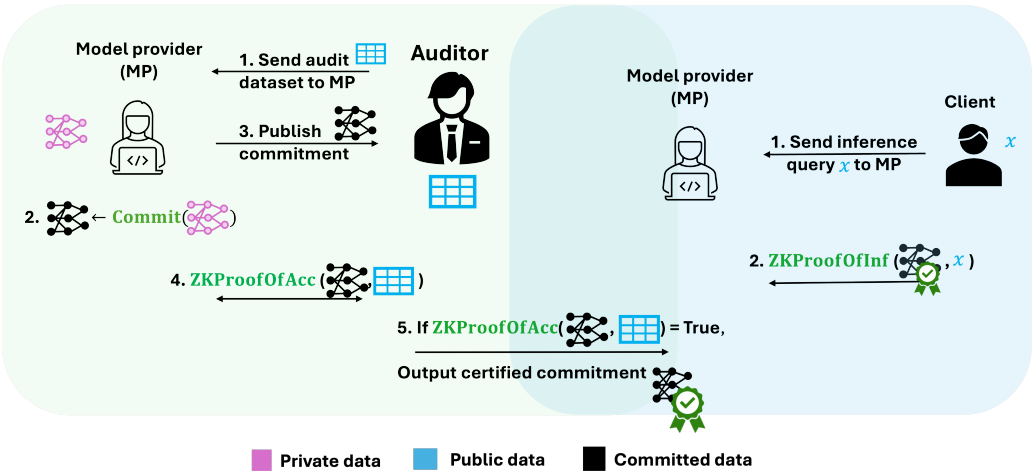

Figure 1: Simplified protocol flow for (insecure) ZK-based ML certification. Left: The model provider, after observing the audit dataset, commits to a model and engages with the auditor in a zero-knowledge proof of accuracy (ZKProofOfAcc). If the audit succeeds, the auditor certifies the committed model. Right: For each new inference query, the model provider interacts with the client in a zero-knowledge proof of inference (ZKProofOfInf) protocol, ensuring that the result is consistent with the previously certified commitment.

relation $\mathcal{R}$: given $\mathsf{x} = (\mathsf{com}_h, \mathsf{com}_S, S_{\text{audit}})$ as a public statement and $\mathsf{w} = (h, \rho_h, \rho_S)$ as private witness, $\mathcal{R}$ outputs 1 iff (1) $f(h, S_{\text{train}}, S_{\text{audit}}) = 1$, (2) h is the ML model (or hypothesis) committed in com_h under randomness ρ_h, and (3) S_{train} is the training data (which can also include randomness used for training if the training algorithm is randomized) committed in com_S under randomness ρ_S. With these building blocks, we define an (insecure) CMC scheme as follows:

1. Auditor samples S_{audit} from some distribution $\mathcal{D}$ (or uses a public one) and sends S_{audit} to the model provider.
2. Model provider sends cryptographic commitments to its model $\mathsf{com}_h = \mathsf{Commit}(h; \rho_h)$ and to the training data $\mathsf{com}_S = \mathsf{Commit}(S_{\text{train}}; \rho_S)$ to the auditor, where ρ_h and ρ_S are sampled uniformly from the randomness space.
3. They interact to execute ZKP: $b \leftarrow \langle \mathcal{P}(h, S_{\text{train}}), \mathcal{V} \rangle (\mathsf{com}_h, \mathsf{com}_S, S_{\text{audit}})$, where model provider plays $\mathcal{P}$ and the auditor plays $\mathcal{V}$ and outputs b.

If the output b is 1, the auditor gets assurance that private model h and training data S_{train} satisfy relation $\mathcal{R}$ without learning anything else about h or S_{train}. We note that, depending on the concrete instantiation of f (see § 3.4 for examples), some steps may be omitted. For example, [56] verifies accuracy of a model h, and their predicate f only needs to check that h's empirical error on S_{audit} is below a given threshold. Hence, their protocol does not require S_{train} and com_S since they are never used. We keep our syntax intentionally general to capture a wide range of prior works.

Security of CMC Schemes At first glance, one might expect the security of the above CMC scheme to follow directly from the aforementioned properties of the underlying zero-knowledge proof system ZKP and commitment scheme Commit. This is indeed the case for confidentiality, i.e., privacy of the model and training data. However, as we show in Section 4, (knowledge) soundness of ZKP and binding of Commit do *not* necessarily imply the desired notion of soundness: a certified model may satisfy the audited property on the specific audit dataset yet fail to generalize beyond it. In particular, the soundness of ZKP is defined only with respect to a fixed statement $\mathsf{x} = (\mathsf{com}_h, \mathsf{com}_S, S_{\text{audit}})$, which binds the proof to a *particular* dataset S_{audit}. Consequently, the proof alone does not ensure good behavior on *fresh* data sampled from the distribution, which is the central goal of auditing. To the best of our knowledge, prior CMC work does not formalize appropriate security goals. In Section 5, we introduce formal definitions that address it and present a CMC scheme that satisfies them.

3.4 Example of Certification Predicates

We now list concrete examples of predicates f considered in prior works on CMC.

Certifying Accuracy To certify accuracy, we consider the empirical error rate as follows:

$$\hat{\ell}_S(h) = \frac{1}{n} \sum_{(x,y) \in S} \mathbb{I}(h(x) \neq y)$$

where $n = |S|$, x is a data sample, and y is the corresponding label. For a public error threshold t, the auditor can check whether the model h has empirical error below t on the audit dataset S_{audit} by defining the corresponding predicate f as follows:

$$f(h, S_{\text{audit}}) = 1 \iff \hat{\ell}_{S_{\text{audit}}}(h) \leq t$$

Certifying Fairness with Demographic Parity Demographic parity [11] is one of the most basic fairness metrics, measuring the difference between the prediction probabilities conditioned on a sensitive attribute. We consider the empirical parity differences as follows:

$$\Delta_{\mathsf{dp}}(h, S_{\text{audit}}) = \left| \frac{1}{n_0} \sum_{x \in S_0} \mathbb{I}(h(x) = 1) - \frac{1}{n_1} \sum_{x \in S_1} \mathbb{I}(h(x) = 1) \right|$$

where s_x denotes the sensitive feature of a data point x, $S_0 = \{x \in S_{\text{audit}} : s_x = 0\}$, $S_1 = \{x \in S_{\text{audit}} : s_x = 1\}$, $n_0 = |S_0|$, and $n_1 = |S_1|$. With this definition, the model is considered "fair" if the parity difference Δ_{dp} is low. For a public parity threshold t, the auditor can check whether the model h has empirical parity difference below t on the audit dataset S_{audit} by defining the corresponding predicate f as follows:

$$f(h, S_{\text{audit}}) = 1 \iff \Delta_{\mathsf{dp}}(h, S_{\text{audit}}) \leq t$$

Certifying Correct Training Procedure To check whether a model h is correctly trained according to a specified training procedure Train on a training dataset S_{train}, we define the predicate f as follows:

$$f(h, S_{\text{train}}) = 1 \iff h = \mathsf{Train}(S_{\text{train}})$$

where a randomized training procedure can be captured by assuming S_{train} contains training randomness.

More Complex Predicates We note that these examples are not exhaustive, and more complex predicates can be constructed by defining an arbitrary predicate f amenable to general-purpose zero-knowledge proofs or by combining multiple predicates. For instance, one may define f to certify both accuracy *and* correct training procedure simultaneously as follows:

$$f(h, S_{\text{train}}, S_{\text{audit}}) = 1 \iff \hat{\ell}_{S_{\text{audit}}}(h) \leq t \wedge h = \mathsf{Train}(S_{\text{train}})$$

4 Attacking CMC

Returning to our example, suppose the insurance company saves costs by *denying* claims. Intuitively, an accuracy audit with provable guarantees—such as those provided by ZK proof-based systems—and with a sufficiently high threshold (e.g., passing only if accuracy on the auditor's dataset exceeds 95%) should prevent the company from deploying a model that unjustifiably denies too many claims.

We show that this intuition is false. Because machine learning is inherently data-dependent, certified properties need not hold once the model is deployed and applied to fresh data, even when drawn from the same distribution. More formally, while prior works certify that

$$f(h, S_{\text{train}}, S_{\text{audit}}) = 1$$

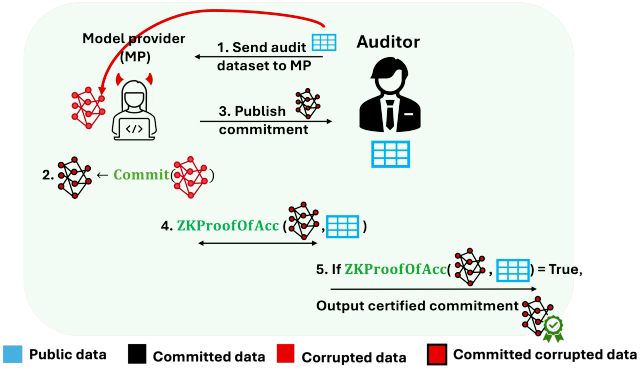

Figure 2: Protocol flow during a data forging attack. The model provider, i.e., the attacker, after observing the audit dataset, commits to a model engineered using the knowledge (indicated by red arrow) of this dataset. Then, it engages with the auditor, i.e., the defender, in a zero-knowledge proof of accuracy (ZKProofOfAcc). If the audit succeeds, the auditor certifies the committed corrupted model.

for some predicate f, a given model h, training data S_{train}, and audit dataset S_{audit}, this does not imply that the stronger property F such that

$$F(h, S_{\text{train}}) = 1 \iff \Pr_{S_{\text{test}} \leftarrow \mathcal{D}}[f(h, S_{\text{train}}, S_{\text{test}}) = 1] > \mathsf{p},$$

where p is a (non-negligible) probability and $\mathcal{D}$ is a distribution over the entire population $Q = \{(x_i, y_i)\}_{i=1}^{m}$. The true goal of an audit, however, is precisely such stronger guarantees: an auditor typically seeks to ensure that a model remains fair, accurate, or robust not only on a particular dataset, but also on the unseen datasets it will encounter during deployment.

We show that this gap can be exploited. In particular, if S_{audit} is partially known to the model provider before it is required to cryptographically commit to the model, the provider can ensure $f(h, S_{\text{train}}, S_{\text{audit}}) = 1$ (and thus pass the audit), *without* additionally satisfying F, which is the actual intended security property (see Fig. 2 for a visualization). A malicious model provider has strong incentives to do so: for example, the insurance company could deploy a model that maximizes accuracy on the audit dataset (and thus passes the audit), yet still unjustifiably denies numerous insurance claims.

Attack Game with Known Audit Data Before providing a concrete attack example, we introduce a theoretical tool – an attack game – which showcases the gap between verifying $f(h, S_{train}, S_{audit})$ (which is what prior approaches certified) and $F(h) = (\Pr_{S_{test} \leftarrow \mathcal{D}}[f(h, S_{train}, S_{test}) = 1] > \mathsf{p})$ (the intuitive property that one would want to ensure) for CMC schemes where the model provider is given the audit dataset at the beginning of the audit process.

For simplicity, we will assume that the audit process verifying $f(h, S_{train}, S_{audit})$ is perfectly secure, i.e., the outcome of $\langle \mathcal{P}(h, S_{\text{train}}), \mathcal{V} \rangle(\mathsf{com}_h, \mathsf{com}_S, S_{\text{audit}})$, where com_S is a commitment to S_{train} and com_h is a commitment to h, is 1 if and only if $f(h, S_{train}, S_{audit}) = 1$.

In the game, the adversary will win only if it can come up with a model h and training data S_{train}, such that: **(1)** $f(h, S_{train}, S_{audit}) = 1$, i.e, the adversary would pass an audit on the dataset S_{audit}, and **(2)** $F(h, S_{train}) = 0$. To make the attack even stronger, we require the adversary to additionally satisfy a utility requirement (formalized via a predicate L) in order to win the game. Intuitively, the goal of L is to capture the actual intent of the malicious model owner: For example, in case of the insurance company that wishes to deny claims, we could use $L(h) = \Pr_{x \sim \{0,1\}^d}[h(x) = 0] > 0.9$.

Definition 3 (Adaptive Training with Known Auditing Data)**.** *Let $\mathcal{M}$ be the model space, $\mathcal{X}$ be the data space and $\mathcal{D}$ be a distribution over $\mathcal{X}$. Let $f : \mathcal{M} \times \mathcal{X}^k \times \mathcal{X}^\ell \to \{0,1\}$ be a predicate verified by the model certification, and let $F : \mathcal{M} \times \mathcal{X}^k \to \{0,1\}$ be the actual intended security property. Let L denote the utility predicate*[8]*. Consider the following game played by an adversary $\mathcal{A}$:*

1. *Sample $S_{audit} \sim \mathcal{D}^\ell$.*
2. *Given S_{audit}, $\mathcal{A}$ outputs a hypothesis h_A and a training dataset S_{train}.*
3. *Obtain $b = f(h_A, S_{train}, S_{audit})$.*
4. *The output of the game is 1 ($\mathcal{A}$ 'wins') iff $b = 1$, $F(h_A, S_{train}) = 0$, and $L(h_A, S_{train}) = 1$.*

   The output is 0 ($\mathcal{A}$ 'loses') otherwise.

Restricting Knowledge of Audit Data Although the attack game in Def. 3 allows an adversary to learn the entire audit dataset S_{audit}, some of our results indicate that the attack can be successful even in a more challenging scenario where the adversary only has *partial* knowledge of the audit dataset. For example, our experimental results in § 7 exhibit that the attacks on accuracy and fairness certification are effective even when the adversary only has access to a small fraction of the audit dataset. Moreover, our attack on differential privacy certification in § 4.1 does not require any knowledge of the audit dataset, as the adversary does not need to forge data based on the audit dataset in order to succeed.

4.1 Example Data Forging Attacks

To demonstrate the breadth of the attack surface for data forging, we now give three concrete examples of data forging attacks: two enable an attacker to evade audits on accuracy (e.g. [56]) and fairness respectively within the framework of Def. 3. The third extends beyond this framework (specifically, it does not require the framework's first step, and the adversary **does not** obtain audit data) to attack differential privacy certification (e.g. [41]). Our example attacks target decision tree (DT) models as a proof of concept for ease of exposition and due to the ubiquity of real-world DT deployments. We show empirically that other model types can be effectively attacked as well in Sections 7 and G.

We first consider our running example of an insurance company audit. Say the company uses a *decision tree* model(see §B.2 for background).The auditor wishes to check that the model is highly accurate, i.e., $F(h) = \Pr_{x \sim \{0,1\}^d}[h(x) = y_x] > 0.95$, where y_x is the true label of x. The insurance company saves on costs when it denies claims, i.e., the model provider's utility $L(h) = (\Pr_{x \sim \{0,1\}^d}[h(x) = 0] > 0.9)$.[9] Following Def. 3, given audit dataset S_{audit}, we wish to construct a hypothesis h_A that passes the audit, while simultaneously achieving $F(h_A, S_{train}) = 0$ and $L(h_A, S_{train}) = 1$.

Attacking Accuracy Certification At a high level, Algorithm 1 constructs a dataset S'_{train} which causes h_A to overfit to S_{audit}, allowing the attacker to achieve distinct levels of performance on the audit data compared to the rest of the distribution. In more detail, we build S'_{train} from the audit dataset S_{audit} using the corresponding true labels—to ensure the model passes the accuracy test on S_{audit}—and augment it with points in the neighborhood of those audit points. The points in the neighborhood are labeled 0, so that outside the audit samples the model learns to label every point with a 0. More formally, to construct S'_{train}, for every point $r \in S_{audit}$ and every (numeric) feature i, we add $r + \varepsilon \vec{b}_i$ with label 0, where ε is some small number and $\vec{b}_i$ is the ith basis vector. We also add every $r \in S_{audit}$ to S'_{train} with honest labels (see Algorithm 1 for formal description). Thus, the time complexity of the attack is $O(|S_{audit}|d)$. After constructing the forged data, we train a tree on S'_{train} via a standard DT training algorithm until leaves are homogeneous.

Consider a baseline CMC protocol for accuracy auditing, wherein a model owner must prove (i) that a decision tree M was trained correctly from a confidential dataset, and (ii) that the predictions M makes on an audit set are accurate. By inputting data from Algorithm 1 and a DT trained on this data to such a CMC protocol, an attacker can manipulate the accuracy of the model on the audit set and the rest of the distribution essentially independently. The CMC will not catch this capability, since the attacker 'honestly' executes both the verified training and verified inferences in the audit. The attack lies not in the process that the adversary used to obtain the model from the data, but rather in the data itself. The malfeasance is *upstream* of anything this baseline CMC protocol certifies.

We demonstrate this idea empirically in Section 7, and formally below. In particular, Theorem 1 shows that Algorithm 1 allows an adversary to win the attack game of Definition 3.

[8] We assume that distribution $\mathcal{D}$ is implicitly "known" to F and L (it is either hard-coded or provided as a parameter to each). For simplicity of notation, we omit $\mathcal{D}$ from the description of F and L.

[9] For simplicity, we consider datapoints in $\{0,1\}^d$

Algorithm 1 Data Forging to Evade Accuracy Audits

Input: Audit set S_{audit}, dimension d, $\varepsilon > 0$
Output: Training data S'_{train}

```
function AttackA(S_audit, d, ε, g)
    S'_train ← S_audit
    for r ∈ S_audit do
        for i ∈ [d] do
            r_0 ← (r + ε b⃗_i, 0)    ▷ b⃗_i is the one-hot vector in dimension i
            r_1 ← (r − ε b⃗_i, 0)
            S'_train ← S'_train ∪ {r_0, r_1}
    return S'_train
```

As we confirm in Fig. 6, this attack achieves good results. We now show that this attack allows the adversary to win in the attack game of Def 3. To this end, we first state the following theorem:

Theorem 1. *Decision tree training until leaves are homogeneous on the output of Algorithm 1 yields a tree $\mathcal{T}$ such that for every $x \in \mathbb{R}^d$, $\mathcal{T}(x) = 1$ only if $||x - r||_\infty < \varepsilon$ for some $r \in S_{audit}$.*

At a high level, the proof shows that if two points land in the same leaf, then any point lying between them on one coordinate must also fall in that leaf. Further, as the only non-zero points in S'_{train} are audit points, every non-zero-labeled leaf contains an audit point. For any x at least ε away from all audit points, if $\mathcal{T}(x) = 1$, one can construct a nearby training point with label 0 that must lie in the same leaf, giving a contradiction as we trained until homogeneity. See §C for details. □

Thus, whenever a model provider generates a training dataset using Algorithm 1, an honestly trained decision tree that grows until homogeneity will achieve perfect accuracy on the audit dataset, yet predict zero for all inputs that lie outside an ε-neighborhood of the audit dataset points. Thus, for an appropriate choice of epsilon (which varies depending on the target distribution), the adversary wins in the game specified in Definition 3 with probability one.[10]

Attacking Fairness Certification A similar method of attack can be used to evade fairness audits as well. Algorithm 2 demonstrates that the model owner can manipulate the sensitive attributes of a dataset to achieve a desired level of fairness (measured via e.g. demographic parity) on audit data.

As in Algorithm 1, this attack enforces certain behavior on the audit set while learning a different function on the rest of the distribution. Instead of returning the true labels on the audit dataset (which encourages the audit to appear highly accurate), we return a random label, causing the model to appear completely fair, as the $k = 0$ and $k = 1$ groups have the same outcome distribution in the audit dataset. Outside of the audit set, however, the model learns to label points according to their sensitive attribute, which results in maximally unfair decisions.

[10] Assuming that a model which almost always outputs 0 is not highly accurate in our scenario.

Algorithm 2 Data Forging to Evade Fairness Audits

Input: Audit set S_{audit}, dimension d, $\varepsilon > 0$, sensitive attribute k
Output: Training data S'_{train}

```
function AttackF(S_audit, d, ε, g)
    S'_train ← ∅
    for x, y ∈ S_audit do
        y' ←$ {0, 1}
        S'_train ← S'_train ∪ (x, y')
    for r ∈ S_audit do
        for i ∈ [d] do
            r_0 ← (r + ε b⃗_i, r[k])
            r_1 ← (r − ε b⃗_i, r[k])
            S'_train ← S'_train ∪ {r_0, r_1}
    return S'_train
```

Attacking Differential Privacy Certification We also demonstrate a data forging attack which subverts zero-knowledge proofs of differential privacy certification. The goal of differential privacy is to protect the privacy of users in a dataset, most commonly by adding a controlled amount of random noise to the outputs [19]. This induces a tradeoff between utility and privacy – tighter protections on privacy require the addition of greater amounts of noise, which degrades the utility of the output. This tradeoff creates a conflict of interest, since institutions are often incentivized to choose high utility over privacy protections [28].

To address this, Shamsabadi et al. describe a scheme to certify the differential privacy of a model [41]. It works by having the model owner perform a zero-knowledge proof, which shows that they correctly performed differentially private stochastic gradient descent (the standard algorithm for training differentially private models [43]) on a pool of committed data. In our syntax, their protocol be captured by instantiating the predicate f for "correct training procedure" (see § 3.4). However, this scheme's privacy analysis is broken if the adversary alters a dataset via Algorithm 3, resulting in a model with dramatically worse privacy than what is claimed by the certificate.

The attack is straightforward: the attacker creates several copies of the data points in S_{train} whose privacy they want to erode (denoted as the target set T), and then uses the modified dataset as an input to the scheme in [41]. In differential privacy, the *sensitivity* of a training algorithm (i.e. how much model parameters can vary if a single input is changed) is used to calibrate the amount of noise which needs to be added

to the model parameters in order to protect the privacy of training data. By copying points in the dataset, Algorithm 3breaks the sensitivity assumptions that differentially private training algorithms rely upon. This is similar to the phenomenon noted in Wilson et al. [50] who observed that standard differential privacy guarantees over SQL queries fail when multiple queries belong to a single user.

Algorithm 3 Data Forging to Weaken Differential Privacy

Input: dataset S_{train}, set of targets $T \subset S_{train}$, duplicity n
Output: Training data S'_{train}

```
function AttackDP(S_train, T, n)
    S'_train ← S_train
    for x ∈ T do
        add n − 1 copies of x to S'_train
    return S'_train
```

Together these three straightforward example attacks demonstrate the breadth of the data forging attack surface. The model owner's full control over data can be used to attack certificates of multiple properties, with and without prior knowledge of audit data.

4.2 Statistical Methods Fail to Detect Data Forging

One might seek to strengthen CMC protocols against attacks in the previous section by leveraging statistical tests, e.g., Welch's t-test to assess whether the model owner's dataset was drawn from a different distribution than the audit set. Indeed previous studies propose zero-knowledge proofs of statistical tests for this purpose [12, 18]. In §Dhowever, we prove that adversaries can forge datasets which successfully circumvent CMCs augmented with a variety of statistical tests.

Our results show that an attacker can forge datasets that successfully evade CMC accuracy audits, while yielding outputs which are indistinguishable from the audit set for a broad family of functions. At a high level, we demonstrate that the attacker can cause the distributional properties of the training data to converge towards the audit data without sacrificing the efficacy of the attack in Algorithm 1 by adding more copies of the audit data to the training data. This causes the audit data and the training data to appear as if they were drawn from the same distribution under a variety of statistical tests without impacting the model's ability to learn the desired behavior.

Thus, functions within this family cannot effectively thwart attackers satisfying Definition 3. We apply this result in Corollary 1 which shows that a data forging attacker can always evade detection by Welch's t-test.

Corollary 1. *Given an audit dataset S_{audit} and significance level α, we can use Algorithm 1 to construct a training dataset S'_{train} such that for any feature j, S'_{train} passes Welch's t-test when its values in feature j are compared to those of S_{audit} with significance level α.*

Proof of Corollary 1 and the results leading up to it are deferred to Appendix Dfor brevity.

The difficulty of detecting data forging attacks against CMCs, combined with the variety of certification protocols they can evade, motivate a new perspective on CMC security which we present in the following section.

5 Secure Cryptographic Model Certification

In this section, we present our positive results on *provably secure* cryptographic model certification (CMC) protocols, with full details provided in § F. We begin by introducing the syntax and security definitions for CMC in § 5.1. We then present a template for secure CMC and our main theorem in § 5.2, followed by an example instantiation for accuracy auditing in § 5.3. As we will later see in the case study, our protocol template allows us to restore/expand the security of some of the constructions from Table 1.

5.1 Syntax and Security Definitions

Definition 4 (Cryptographic Model Certification Protocol)**.** *A cryptographic model certification (CMC) protocol Π for a distributional predicate F is a tuple of algorithms* $(\mathsf{Commit}, \mathsf{Prove}, \mathsf{Audit})$*: a commitment, a proving, and an auditing algorithm. Let* $(\mathsf{com}, b) \leftarrow \langle \mathsf{Prove}(h, S_{train}), \mathsf{Audit} \rangle$ *denote the interaction between* Prove *and* Audit*, where* Prove *takes a hypothesis h and optionally a training dataset S_{train} as private input and outputs a commitment* com *during an execution, and* Audit *halts by outputting a binary b.*

We propose the following security properties for a CMC protocol Π for a distributional predicate F. We focus on high-level intuition and defer formal cryptographic definitions to § F.1.While completeness, binding, and zero knowledge are direct adaptations of those of commitments and zero-knowledge proofs(see § B.3 and § B.4), $\tilde{F}$-relaxed knowledge soundness is a novel property tailored to our CMC setting. In particular, we call this property "$\tilde{F}$-relaxed" to allow parameterizing the soundness requirement by a predicate $\tilde{F}$ that is only a close approximation of the desired property F. As we shall see in concrete examples, this relaxation turns out necessary because the auditor only checks the property f on a *finite sample*, which may not perfectly reflect the property F on the *underlying distribution*. In our concrete instantiations, we will quantify the gap between F and $\tilde{F}$(see § F.2.1 and § F.2.2).

Completeness. If an honest prover holds a model h and a training set S_{train} such that $F(h, S_{\text{train}}) = 1$, then this prover should pass the audit (i.e, Audit outputs $b = 1$) except with small error probability p.

Binding. No computationally bounded prover can change its model h or S_{train} after generating a commitment com to them.

Zero Knowledge. An honest execution of the protocol between Prove and Audit does not reveal any information about the model h beyond the fact that $F(h, S_{\text{train}}) = 1$.

$\tilde{F}$-relaxed Knowledge Soundness. If a (potentially dishonest prover) Prove^* commits to a pair of invalid model h and training dataset S_{train}, i.e., $\tilde{F}(h, S_{\text{train}}) \neq 1$, then Audit should detect it by outputting $b = 0$ with high probability. In slightly more detail, for any Prove^*, there exists an efficient *knowledge extractor* Ext that extracts a pair (h, S_{train}) from Prove^* such that the above condition holds (except with small error probability κ).

5.2 Secure CMC Template

We now define a *commit-sample-prove* CMC protocol $\Pi_{\text{csp}} = (\mathsf{Commit}, \mathsf{Prove}, \mathsf{Audit})$. The protocol starts with the model provider committing to its model h using a commitment scheme Commit(§ B.4).Upon receiving the commitment, the auditor responds with an *audit dataset* S_{audit}, consisting of n i.i.d. samples drawn from a distribution $\mathcal{D}$ over the query space $Q = \{(x_i, y_i)\}_{i=1}^m$. Finally, the model provider and the auditor run a zero-knowledge proof (ZKP) protocol $\mathsf{ZKP} = (\mathcal{P}, \mathcal{V})$ to prove that the committed model h satisfies an *empirical predicate* f on the sampled audit dataset S_{audit}, i.e., ZKP for the relation $\mathcal{R}$ such that $(\mathsf{x}, \mathsf{w}) \in \mathcal{R} \iff f(h, S_{\text{audit}}) = 1 \wedge \mathsf{com} = \mathsf{Commit}(h; \rho)$. Here, the empirical predicate f should be chosen to closely approximate the desired property F over the underlying distribution $\mathcal{D}$; we will formalize this requirement as small *false negative rate*. Moreover, f should be chosen such that the *false positive rate* of f is small with respect to the relaxed predicate $\tilde{F}$. Intuitively, this means that if a model h does not satisfy the relaxed property $\tilde{F}$, then it should fail the audit with high probability (i.e., soundness), while if a model h satisfies the desired property F, then it should pass the audit with high probability (i.e., completeness). While we focus on a protocol checking F on a hypothesis h only, the construction below can be naturally extended to a more complex F and f that additionally take a training dataset S_{train} as input.

$\langle \mathsf{Prove}(h), \mathsf{Audit} \rangle$

1. Prove computes $\mathsf{com} = \mathsf{Commit}(h; \rho)$ using a uniformly random string ρ and sends com to Audit.
2. Audit samples $S_{\text{audit}} \sim \mathcal{D}^n$ and sends it to Prove.
3. Prove and Audit execute $b \leftarrow \langle \mathcal{P}(\mathsf{w}), \mathcal{V} \rangle(\mathsf{x})$, where $\mathsf{x} = (\mathsf{com}, S_{\text{audit}})$ and $\mathsf{w} = (h, \rho)$. Here, Prove plays $\mathcal{P}$ and Audit plays $\mathcal{V}$ of ZKP.
4. Prove outputs com, while Audit outputs b.

The key takeaway is that the audit dataset S_{audit} should be chosen *independently* of the model h and the training dataset S_{train}. Sampling S_{audit} after h is committed ensures independence, rendering our earlier attack ineffective. The following theorem (formally stated and proved in § F.2) states that, if the empirical predicate f satisfies the bounds mentioned below, Π_{csp} satisfies the desired security properties in a general fashion. **Essentially, our result allows the protocol designer to choose an empirical predicate f that approximates the desired property F well enough, and then plug in any commitment scheme (§ B.4)and ZK proof system (§ B.3)satisfying the required security properties (§ B.3,§ B.4)to get a secure auditing scheme.**

Theorem 3 (informal). *Let the empirical predicate f, the distributional predicate F, and the relaxed distributional predicate $\tilde{F}$ satisfy the following false negative and false positive rate bounds for every model h:*

$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) \neq 1 \mid F(h) = 1] \leq p_{fnr}$$
$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) = 1 \mid \tilde{F}(h) \neq 1] \leq p_{fpr}$$

If Commit *and* ZKP *satisfy the standard security properties (§ B.3-B.4), then the protocol Π_{csp} is **complete** with error probability p_{fnr}, **binding**, **zero knowledge**, and **$\tilde{F}$-relaxed knowledge sound** with error probability $\approx p_{fpr}$.*

Proof sketch. We focus on $\tilde{F}$-relaxed knowledge soundness, as the other properties directly follow from the underlying commitment scheme and ZK proof system with standard security guarantees. Our proof carefully unifies standard arguments from learning theory with the proof technique of "knowledge extraction with interleaved probabilistic tests", commonly used in advanced proof systems [4, 10, 37]. To illustrate the intuition, let us assume the commitment scheme is perfectly binding and straightline extractable, i.e., once the prover sends com to the auditor, one can immediately extract a unique model h and a unique randomness ρ such that $\mathsf{com} = \mathsf{Commit}(h; \rho)$. Such a commitment scheme can be constructed in the common reference string (CRS) model using public key encryption. If the committed model h does *not* satisfy $\tilde{F}$, then by the assumption on the false positive rate, the probability that $f(h, S_{\text{audit}}) = 1$ is at most p_{fpr} over the choice of S_{audit}.

The actual proof is significantly more involved, when the commitment scheme is *only computationally binding* (which is the case for most practical instantiations). In particular, we need to construct a meta-extractor that runs the knowledge extractor $\mathcal{E}$ for ZKP to obtain a candidate model h, and rewinds $\mathcal{E}$ with fresh $S'_{\text{audit}} \sim \mathcal{D}^n$ to ensure the validity of h via probabilistic tests (or otherwise break binding of the commitment). Our formal security proof takes care of these subtle technicalities.

5.3 Example: CMC for Accuracy

We detail example instantiations of f and F for accuracy (§ F.2.1)and demographic parity auditing(§ F.2.2), both of which enable arbitrary small false positive and negative rates with a sufficient number of samples n. To instantiate Π_{csp} for accuracy auditing, we consider the empirical and true error rates as follows:

$$\hat{\ell}_S(h) = \frac{1}{n} \sum_{(x,y)\in S} \mathbb{I}(h(x) \neq y) \quad \ell(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(h(x) \neq y)]$$

where $n = |S|$, and define the empirical predicate f, the distributional predicate F, and the relaxed predicate $\tilde{F}$ as follows.

$$\begin{aligned} f(h, S_{\text{audit}}) = 1 &\iff \hat{\ell}_S(h) \leq t + \delta \\ F(h) = 1 &\iff \ell(h) \leq t \\ \tilde{F}(h) = 1 &\iff \ell(h) \leq t + 2\delta \end{aligned}$$

With these definitions in place, we can bound both the false negative p_{fnr} and false positive rates p_{fpr} by $2e^{-2n\delta^2}$ by a standard concentration bound argument (see Lemma 3 in § F.2.1).By plugging them into Theorem 3, we conclude that Π_{csp} instantiated for accuracy auditing is a secure CMC with the following properties for a sufficiently large $n\delta^2$.

- If the true error of the model h is $\leq t$, then the auditor accepts with high probability.
- If the auditor accepts, then the auditor gets assurance that the true error of the model h is at most $t + 2\delta$, even if the prover is misbehaving.

As usual in learning theory, there is a trade-off between the sample size n and the tightness of the confidence parameter δ. By tuning δ appropriately, one can reduce the number of samples n needed to achieve desired false positive and negative rates. This is particularly important in practice since 1) the computational cost of ZKP typically scales with n, and 2) gathering large audit datasets can be challenging. For instance, by choosing $(n, \delta) = (10\text{k}, 0.02)$ or $(n, \delta) = (40\text{k}, 0.01)$, we can achieve both false positive and negative rates ≤ 0.001. Since typical dataset sizes in practice are in the order of tens or hundreds of thousands, these configurations enable high-confidence auditing while maintaining reasonable computational costs for ZKP.

6 Case Studies: Vulnerability to Data Forging in Previous Work

Our formalization from §5 and §4 lets us test whether a protocol is vulnerable to data-forging, and provides a way to lift the security guarantees of prior works to resist data forging attacks. For CMC works which reveal neither the model nor the training data, the most important check boils down to whether the prover is required to commit to the training data and/or to the model before seeing the audit dataset.

We examined several prior works [9,21,33,35,41,42,48,51,56] with formal security guarantees. Surprisingly, the majority of them either do not explicitly state when the audit dataset is revealed, or assume the prover's training dataset and/or the model itself to be trusted (and are susceptible to data forging if the prover is actually malicious). Works that do not discuss the timing of the commitment often point out that their solution can be used to conduct audits using publicly known datasets, in which case the public dataset can be assumed to be known to the adversary prior to auditing, making the solution vulnerable to data-forging.

We present two of our case studies below, and discuss the rest in §E. We note that works that are not susceptible to our data-forging attacks nonetheless provide only dataset-specific guarantees, i.e., their proofs certify properties solely on the chosen audit set/inference queries already submitted by clients, without extending to the underlying data distribution. It would be interesting to perform an analysis similar to that in §F.2.1 and § F.2.2 to derive formal guarantees that hold for the corresponding distributions. We summarize the results of our findings in Table 1.

6.1 Zero Knowledge Proofs for Decision Tree Predictions and Accuracy

Goal and Solution Details. Zhang et al. [56] introduce protocols for auditing accuracy and verifying decision tree predictions without revealing any additional information about the model itself. Zhang et al.'s main contribution is in designing a custom zero-knowledge proof tailored to efficiently verifying the decision tree prediction. The proof consists of algorithms to generate public parameters, custom commitment algorithm for decision tree models, the prover's algorithm which outputs a proof of inference/accuracy, and verifier algorithm to check this proof. The prover, i.e., model provider, must first commit to its model and subsequently demonstrate that the predictions on client queries are consistent with this commitment. For accuracy verification, the authors propose a batching technique that allows to more efficiently checks the correctness of model predictions across multiple inputs, followed by a ZKP-verified calculation of the proportion of predictions that match the ground truth labels of the data.

Security Model. Zhang et al.'s security notion follows the traditional definition structure of a zero-knowledge proof system, where the protocol is required to satisfy correctness, soundness, and zero-knowledge. Both parties can be malicious. In the context of our analysis we are interested in soundness, which specifies whether a malicious prover can deceive the verifier (i.e., auditor), that the prover's hypothesis passed the test. At a high level, the authors' soundness definition can be summarized as follows: A prover should not be able to output a commitment to a tree $\mathcal{T}$ along with a proof π, an

accuracy threshold *acc*, and a dataset D, such that the verifier accepts the proof, and at the same time, $\mathcal{T}'s$ accuracy on D is not equal to *acc*.

Discussion. The security notion that Zhang et al's work provides aligns well with the intuitive goals of verifying both the correctness of individual predictions and the accuracy of a model on a given dataset. However, it does not provide formal guarantees for datasets beyond the audited one, i.e., the accuracy verification solution does not generalize to other datasets drawn from the same distribution. The protocol description further does not explain how the audit dataset is sampled. In fact, Zhang et al. explicitly note that it is possible to use their solution to check accuracy on a *public dataset*. In this setting, their approach falls within our framework of Definition 3, and is vulnerable to the same attack as outlined in §4.1. In fact, note that our example works even if the prover supplies an *additional* proof of training to complement its proof of accuracy.

Expanding/Restoring Security. Using our techniques, Zhang et al.'s construction can be easily adapted to satisfy the guarantee with respect to new datasets drawn from the same distribution. In fact, this corresponds precisely to our example in § 5.3, where Zhang et al.'s construction can be plugged in as the ZKP of our protocol Π_{csp} in §5.2.

6.2 Confidential-DPproof: Confidential Proof Of Differentially Private Training

Goal and Solution Details. Shamsabadi et al. [41] present Confidential-DPproof, a framework that enables the model provider to prove to an auditor that their model was correctly trained via DP-SGD [1], a classic approach for training models with differential privacy guarantees. The certification of DP-SGD's training run is done in zero-knowledge.

Security Model. Shamsabadi et al. [41] consider a setting with a mutually distrusting model provider and auditor. The model provider is fully malicious, while the auditor is semi-honest. Confidential-DPproof considers standard definitions of correctness, soundness, and zero-knowledge.

Discussion. As we describe in §3 ("Attacking Differential Privacy Certification"), Shamsabadi et al.'s solution is susceptible to data forging attacks.

Limitations of Our Secure Template in the Privacy Setting. Defining practical empirical predicates for privacy testing is more challenging for privacy than for accuracy or fairness. Empirical audits of differential privacy, such as membership inference attack success rate [31], often incur hefty computational costs. Further, both privacy audits and differentially private training necessitate assumptions about training data which are difficult to validate in the CMC model without further cryptographic infrastructure.

As a result, the approach we introduced above for Zhang et al. [56] does not apply to Confidential-DPproof. The core of the issue that Confidential-DPproof is essentially a proof of training, and the data for this training is necessarily chosen by the model provider, not by the auditor. One possible way to still make use of Confidential-DPproof guarantees would be to verify that the training data was generated by a set of trusted parties independent from the model provider. For instance, each individual in the dataset could sign their data record with a distinct key pair. The service provider could then prove that each training record verifies under a single individual's public key. This would thwart the attack in Algorithm 3, as duplication of data points would not be possible. However, instantiating such a scheme efficiently and proving its security to more general data forging attacks is a challenge beyond the scope of the present paper. We reserve it as an interesting direction for future work.

7 Evaluation

In this section we underscore the importance of data forging attacks by mounting proof of concept attacks for models trained on a variety of datasets. We show that our attack is effective in making inaccurate models appear accurate and unfair models appear fair, and empirically demonstrate a variety of other qualities, e.g. undetectability with a variety of statistical tools. We provide source code for our implementation and for reproducing our experiments.[11]

Experimental Setup. We use six well-known fairness benchmarking datasets in our experiments: ACSEmployment [17], Adult [7], COMPAS [3], German Credit [30], Default Credit [54], and Communities & Crime [39]. We implemented a modified version of our attack from §4.1 in Python 3.12.3 using SciKit-Learn version 1.6.1 that attempts to minimize an objective when deployed and evaluated its performance against a variety of datasets. For a given run, we split the dataset into an evaluation dataset consisting of 30% of the data, an audit dataset containing 1000 data points (500 for German Credit due to its smaller size), and an initial training data set. We represent the interpolation between a fully honest training run and a fully malicious one by the *attack parameter* $p \in [0,1]$. Concretely, p controls two components of the forged training set S'_{train}: (i) the fraction of audit points that contribute $\pm\varepsilon$-perturbed neighbors (with flipped labels), and (ii) the fraction of S_{train} rows whose labels are flipped. At $p = 0$ the construction degenerates to honest training (no perturbed neighbors, no label flips), i.e., $S'_{train} = S_{train}$; at $p = 1$ S_{audit} gets included in S'_{train}, every audit point generates $2d$ flipped-label neighbors, and every S_{train} label is inverted.The specifics of how honest and malicious data points are labeled depends on the objective.

To attack accuracy, we constructed a training dataset using a modified Algorithm 1, labeling the additional data from the attack with $1 - r_y$ rather than 0 and adding them to the initial training data set. For our fairness attack, we constructed the

[11] https://zenodo.org/records/20337318

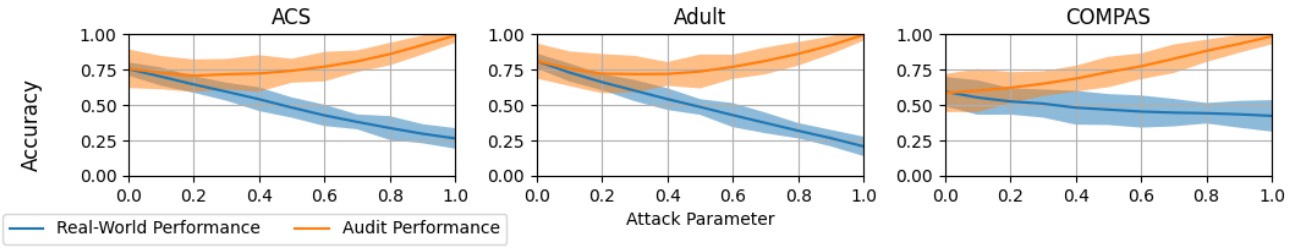

Figure 3: Accuracy of models trained on datasets constructed to minimize real-world accuracy while maintaining high accuracy on audit data for several benchmarks. Attack parameter controls what portion of audit data is known to the model owner and what portion of training data is labeled maliciously. Values are averages over ten runs, error bars represent one standard deviation.

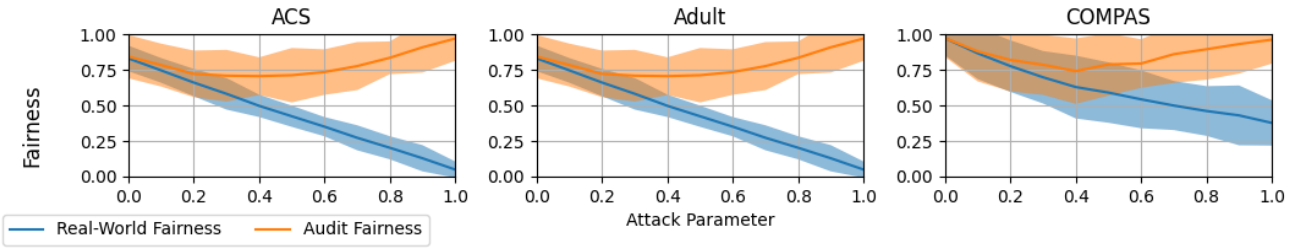

Figure 4: Fairness of models trained on datasets constructed to maintain demographic parity on audit data while producing unfair results on real-world data using various benchmarks. Attack parameter controls what portion of audit data is known to the model owner and what portion of training data is labeled maliciously. Values are averages over ten runs, error bars represent one standard deviation. Fairness means 1 – demographic parity gap.

training data similarly, changing how honest and malicious data points were labeled to match Algorithm 2. Honest data points were given random labels, while malicious data points were labeled according to their sensitive attribute. Both attacks then fit a decision tree to their constructed training data using SciKit-Learn's decision tree classifier class.

To ensure that our attack would evade statistical detection, we added extra copies of the audit data to the training data, as computed in Corollary 2 to pass Welch's t-test with significance level 0.05.

Attacking Accuracy Audits. We ran our attack on six benchmark datasets – three shown in Figure 3 (remaining in Figure 5). Across all datasets, our data forging attack enforces high audit accuracy while simultaneously encouraging low performance on real-world evaluation data. Thus **our attack successfully makes inaccurate models *appear* accurate to an auditor.** In practice, this could mean that a fraudulent insurance provider could enrich themselves by using a highly inaccurate model to deny claims, despite passing accuracy audits.

Attacking Fairness Audits. We also performed the attack while targeting demographic parity (using sex as the sensitive attribute) on three datasets, which we present in Figure 4. We were able to reliably train a model with close to 0 fairness gap on the audit dataset, but close to 1 fairness gap when deployed. In other words, **our attack successfully makes unfair models *appear* fair to an auditor.**

Evading Detection via Statistical Methods. We show how our attack can be executed in ways that evade detection by a variety of statistical approaches in Appendix Table 2. We were able to construct malicious training datasets with summary statistics that match those of the audit dataset very closely, and Welch's t-test and Levene's test regularly concluded that the audit and test datasets were drawn from the same distribution. This is consistent with our theoretical results in Appendix D.

Additional Results. An adversary can use data forging attacks to achieve concrete goals beyond degradation of accuracy or fairness, as we show in the Appendix G. For example, Figure 6 shows how an insurance provider could use our attack to hide the claim denial rate of a model from auditors. We have also conducted preliminary experiments that suggest that variations of our attack generalize to XGBoost (§ G.1) and neural network models (§ G.2).

8 Conclusion and Future Work

This work brings attention to data-dependent vulnerabilities in CMC. We showed that, in several proposed certification frameworks, a provider can pass an audit while undermining its intended guarantees in real-world deployment. We proposed concrete attack strategies and provided preliminary evidence that standard statistical tests are unlikely to reliably

detect them. Empirically, we demonstrated the effectiveness of these attacks for certifications for accuracy and fairness. We then introduced new formal security definitions which allow property certification on data *distributions* (rather than data *sets*), thereby ruling out the data-dependent attacks identified in this work. We then presented a general-purpose protocol satisfying our definition.

Our results should not be interpreted as showing that prior CMC constructions are technically incorrect. Rather, they show that the guarantees they provide may not align with the guarantees required in realistic auditing scenarios. More broadly, our work illustrates the importance of close communication between the cryptography, security, and machine learning communities when designing future certification frameworks, so that formal guarantees are aligned with the threats that arise in deployment.

The attack strategy presented in this work poses several open questions. While we demonstrate the data forging attack is undetected even in the presence of Welch's t-test and Levene's test, it remains to be seen whether other statistical tests could effectively detect the attack. Based on the results that we have derived, we find it unlikely that other statistical tests will be effective in detecting the attack. However, we reserve such analysis for future work. We provide rigorous formal proofs that our attacks are effective on decision trees, and preliminary evidence that a similar approach generalizes to neural networks. Characterizing a formal relationship between neural network model capacity and attack effectiveness could be a promising direction in future work.

Our secure CMC template underscores the importance of keeping audit data hidden until the service provider's model is committed. This imposes a limitation on auditing in practice: auditors must either regularly gather fresh data (since the audit dataset is typically revealed during the audit), use additional cryptographic techniques such as secure multiparty computation to keep data hidden during the audit, or perform continuous auditing on user data. Each of these options has strengths and drawbacks which should be evaluated in more detail by future work. Moreover, while we focused on a CMC template based on fresh audit datasets, it remains unclear how to securely certify distributional properties that do not involve an audit dataset, such as differentially private training. A promising direction for future work is a countermeasure against the third attack in §4.1, which likely requires a more advanced CMC mechanism to certify the well-formedness of the training data.

Acknowledgments & Disclaimer

We thank Chen-Da Liu-Zhang for many helpful discussions, and Jonas Guan, Sierra Wyllie and Mohammad Yaghini for helpful feedback on an earlier draft. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. The work was further supported by the CBI postdoctoral fellowship.

This paper was prepared in part for information purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Kasra Abbaszadeh, Christodoulos Pappas, Jonathan Katz, and Dimitrios Papadopoulos. Zero-knowledge proofs of training for deep neural networks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4316–4330, 2024.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 2016. Retrieved January 15, 2022.

[4] Thomas Attema, Michael Klooß, Russell W. F. Lai, and Pavlo Yatsyna. Adaptive special soundness: Improved knowledge extraction by adaptive useful challenge sampling. Cryptology ePrint Archive, Paper 2024/2038, 2024. URL: https://eprint.iacr.org/2024/2038.

[5] Akhil Bandarupalli, Xiaoyu Ji, Aniket Kate, Chen-Da Liu-Zhang, Daniel Pöllmann, and Yifan Song. Velox: Scalable fair asynchronous MPC from lightweight cryptography. In Chun-Ying Huang, Jyh-Cheng Chen, Shiuh-Pyng Shieh, David Lie, and Véronique Cortier, editors, *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS 2025, Taipei, Taiwan, October 13-17, 2025*, pages 1799–1813. ACM, 2025. doi:10.1145/3719027.3765114.

[6] Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. The security of machine learning. *Mach. Learn.*, 81(2):121–148, 2010.

[7] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[8] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation (extended abstract). In Janos Simon, editor, *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2-4, 1988, Chicago, Illinois, USA*, pages 1–10. ACM, 1988. doi:10.1145/62212.62213.

[9] Jade Garcia Bourrée, Hadrien Lautraite, Sébastien Gambs, Gilles Trédan, Erwan Le Merrer, and Benoît Rottembourg. P2NIA: privacy-preserving non-iterative auditing. *CoRR*, abs/2504.00874, 2025. URL: https://doi.org/10.48550/arXiv.2504.00874, arXiv:2504.00874, doi:10.48550/ARXIV.2504.00874.

[10] Benedikt Bünz and Ben Fisch. Multilinear schwartz-zippel mod N and lattice-based succinct arguments. In Guy N. Rothblum and Hoeteck Wee, editors, *Theory of Cryptography - 21st International Conference, TCC 2023, Taipei, Taiwan, November 29 - December 2, 2023, Proceedings, Part III*, volume 14371 of *Lecture Notes in Computer Science*, pages 394–423. Springer, 2023. doi:10.1007/978-3-031-48621-0_14.

[11] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In Yücel Saygin, Jeffrey Xu Yu, Hillol Kargupta, Wei Wang, Sanjay Ranka, Philip S. Yu, and Xindong Wu, editors, *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009*, pages 13–18. IEEE Computer Society, 2009. doi:10.1109/ICDMW.2009.83.

[12] Ian Chang, Katerina Sotiraki, Weikeng Chen, Murat Kantarcioglu, and Raluca Popa. {HOLMES}: Efficient distribution testing for secure collaborative learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 4823–4840, 2023.

[13] David Chaum, Claude Crépeau, and Ivan Damgård. Multiparty unconditionally secure protocols (abstract). In Carl Pomerance, editor, *Advances in Cryptology - CRYPTO '87, A Conference on the Theory and Applications of Cryptographic Techniques, Santa Barbara, California, USA, August 16-20, 1987, Proceedings*, volume 293 of *Lecture Notes in Computer Science*, page 462. Springer, 1987. doi:10.1007/3-540-48184-2_43.

[14] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016. doi:10.1145/2939672.2939785.

[15] Arka Rai Choudhuri, Aarushi Goel, Matthew Green, Abhishek Jain, and Gabriel Kaptchuk. Fluid MPC: secure multiparty computation with dynamic participants. In Tal Malkin and Chris Peikert, editors, *Advances in Cryptology - CRYPTO 2021 - 41st Annual International Cryptology Conference, CRYPTO 2021, Virtual Event, August 16-20, 2021, Proceedings, Part II*, volume 12826 of *Lecture Notes in Computer Science*, pages 94–123. Springer, 2021. doi:10.1007/978-3-030-84245-1_4.

[16] Ivan Damgård and Jesper Buus Nielsen. Scalable and unconditionally secure multiparty computation. In Alfred Menezes, editor, *Advances in Cryptology - CRYPTO 2007, 27th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2007, Proceedings*, volume 4622 of *Lecture Notes in Computer Science*, pages 572–590. Springer, 2007. doi:10.1007/978-3-540-74143-5_32.

[17] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] Vasisht Duddu, Anudeep Das, Nora Khayata, Hossein Yalame, Thomas Schneider, and N Asokan. Attesting distributional properties of training data for machine learning. In *European Symposium on Research in Computer Security*, pages 3–23. Springer, 2024.

[19] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

[20] Daniel Escudero, Elisaweta Masserova, and Antigoni Polychroniadou. Towards scalable YOSO MPC via packed secret-sharing. In Goichiro Hanaoka and Bo-Yin Yang, editors, *Advances in Cryptology - ASIACRYPT 2025 - 31st International Conference on the Theory and Application of Cryptology and Information Security, Melbourne, VIC, Australia, December 8-12, 2025, Proceedings, Part V*, volume 16249 of *Lecture Notes in Computer Science*, pages 68–101. Springer, 2025. doi:10.1007/978-981-95-5116-3_3.

[21] Olive Franzese, Ali Shahin Shamsabadi, and Hamed Haddadi. Oath: Efficient and flexible zero-knowledge

proofs of end-to-end ml fairness. *arXiv preprint arXiv:2410.02777*, 2024.

[22] Sanjam Garg, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Guru-Vamsi Policharla, and Mingyuan Wang. Experimenting with zero-knowledge proofs of training. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1880–1894, 2023.

[23] Sanjam Garg, Pratyay Mukherjee, Omkant Pandey, and Antigoni Polychroniadou. The exact round complexity of secure computation. In Marc Fischlin and Jean-Sébastien Coron, editors, *Advances in Cryptology - EUROCRYPT 2016 - 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part II*, volume 9666 of *Lecture Notes in Computer Science*, pages 448–476. Springer, 2016. doi:10.1007/978-3-662-49896-5_16.

[24] Craig Gentry, Shai Halevi, Hugo Krawczyk, Bernardo Magri, Jesper Buus Nielsen, Tal Rabin, and Sophia Yakoubov. YOSO: you only speak once - secure MPC with stateless ephemeral roles. In Tal Malkin and Chris Peikert, editors, *Advances in Cryptology - CRYPTO 2021 - 41st Annual International Cryptology Conference, CRYPTO 2021, Virtual Event, August 16-20, 2021, Proceedings, Part II*, volume 12826 of *Lecture Notes in Computer Science*, pages 64–93. Springer, 2021. doi:10.1007/978-3-030-84245-1_3.

[25] Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game or A completeness theorem for protocols with honest majority. In Alfred V. Aho, editor, *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, New York, USA*, pages 218–229. ACM, 1987. doi:10.1145/28395.28420.

[26] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof-systems (extended abstract). In Robert Sedgewick, editor, *Proceedings of the 17th Annual ACM Symposium on Theory of Computing, May 6-8, 1985, Providence, Rhode Island, USA*, pages 291–304. ACM, 1985. doi:10.1145/22145.22178.

[27] Vipul Goyal, Elisaweta Masserova, Bryan Parno, and Yifan Song. Blockchains enable non-interactive MPC. In Kobbi Nissim and Brent Waters, editors, *Theory of Cryptography - 19th International Conference, TCC 2021, Raleigh, NC, USA, November 8-11, 2021, Proceedings, Part II*, volume 13043 of *Lecture Notes in Computer Science*, pages 162–193. Springer, 2021. doi:10.1007/978-3-030-90453-1_6.

[28] Andy Greenberg. How one of apple's key privacy safeguards falls short. *Wired*. URL: https://www.wired.com/story/apple-differential-privacy-shortcomings/.

[29] Moritz Hardt. The emerging science of machine learning benchmarks. *Manuscript. https://mlbenchmarks.org*, 2025.

[30] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[31] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[32] Yuval Ishai, Manoj Prabhakaran, and Amit Sahai. Founding cryptography on oblivious transfer - efficiently. In David A. Wagner, editor, *Advances in Cryptology - CRYPTO 2008, 28th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2008. Proceedings*, volume 5157 of *Lecture Notes in Computer Science*, pages 572–591. Springer, 2008. doi:10.1007/978-3-540-85174-5_32.

[33] Daniel Kang, Tatsunori Hashimoto, Ion Stoica, and Yi Sun. Scaling up trustless DNN inference with zero-knowledge proofs. *CoRR*, abs/2210.08674, 2022. URL: https://doi.org/10.48550/arXiv.2210.08674, arXiv:2210.08674, doi:10.48550/ARXIV.2210.08674.

[34] Benjamin Kreuter, Abhi Shelat, and Chih-Hao Shen. Billion-gate secure computation with malicious adversaries. In Tadayoshi Kohno, editor, *Proceedings of the 21th USENIX Security Symposium, Bellevue, WA, USA, August 8-10, 2012*, pages 285–300. USENIX Association, 2012. URL: https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/kreuter.

[35] Tianyi Liu, Xiang Xie, and Yupeng Zhang. zkcnn: Zero knowledge proofs for convolutional neural network predictions and accuracy. In Yongdae Kim, Jong Kim, Giovanni Vigna, and Elaine Shi, editors, *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 2968–2985. ACM, 2021. doi:10.1145/3460120.3485379.

[36] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[37] Ngoc Khanh Nguyen. *Lattice-Based Zero-Knowledge Proofs Under a Few Dozen Kilobytes*. PhD thesis, ETH Zurich, 10 2022. https://www.research-collection.ethz.ch/handle/20.500.11850/574844.

[38] Christodoulos Pappas and Dimitrios Papadopoulos. Sparrow: Space-efficient zksnark for data-parallel circuits and applications to zero-knowledge decision trees. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3110–3124, 2024.

[39] Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2009.

[40] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[41] Ali Shahin Shamsabadi, Gefei Tan, Tudor Ioan Cebere, Aurélien Bellet, Hamed Haddadi, Nicolas Papernot, Xiao Wang, and Adrian Weller. Confidential-dpproof: Confidential proof of differentially private training. In *International Conference on Learning Representations (ICLR)*, 2024.

[42] Ali Shahin Shamsabadi, Sierra Calanda Wyllie, Nicholas Franzese, Natalie Dullerud, Sébastien Gambs, Nicolas Papernot, Xiao Wang, and Adrian Weller. Confidential-profitt: confidential proof of fair training of trees. In *The Eleventh International Conference on Learning Representations*, 2022.

[43] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

[44] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *NIPS*, pages 3517–3529, 2017.

[45] Haochen Sun, Tonghe Bai, Jason Li, and Hongyang Zhang. Zkdl: Efficient zero-knowledge proofs of deep learning training. *IEEE Transactions on Information Forensics and Security*, 2024.

[46] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *AIES*, pages 303–310. ACM, 2018.

[47] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[48] Haodi Wang and Thang Hoang. ezdps: An efficient and zero-knowledge machine learning inference pipeline. *Proc. Priv. Enhancing Technol.*, 2023(2):430–448, 2023. URL: https://doi.org/10.56553/popets-2023-0061, doi:10.56553/POPETS-2023-0061.

[49] Bernard L Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[50] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private sql with bounded user contribution. *Proceedings on Privacy Enhancing Technologies*, 2020.

[51] Chhavi Yadav, Amrita Roy Chowdhury, Dan Boneh, and Kamalika Chaudhuri. Fairproof: Confidential and certifiable fairness for neural networks. *arXiv preprint arXiv:2402.12572*, 2024.

[52] Andrew Chi-Chih Yao. Protocols for secure computations (extended abstract). In *23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 3-5 November 1982*, pages 160–164. IEEE Computer Society, 1982. doi:10.1109/SFCS.1982.38.

[53] Andrew Chi-Chih Yao. How to generate and exchange secrets (extended abstract). In *27th Annual Symposium on Foundations of Computer Science, Toronto, Canada, 27-29 October 1986*, pages 162–167. IEEE Computer Society, 1986. doi:10.1109/SFCS.1986.25.

[54] I-Cheng Yeh and Che hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473–2480, 2009. URL: https://www.sciencedirect.com/science/article/pii/S0957417407006719, doi:10.1016/j.eswa.2007.12.020.

[55] Guanhua Zhang, Florian E Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. *arXiv preprint arXiv:2506.07673*, 2025.

[56] Jiaheng Zhang, Zhiyong Fang, Yupeng Zhang, and Dawn Song. Zero knowledge proofs for decision tree predictions and accuracy. In *CCS*, pages 2039–2053. ACM, 2020.

[57] Tianyu Zhang, Shen Dong, Oyku Deniz Kose, Yanning Shen, and Yupeng Zhang. Fairzk: A scalable system to prove machine learning fairness in zero-knowledge. In Marina Blanton, William Enck, and Cristina Nita-Rotaru, editors, *IEEE Symposium on Security and Privacy, SP 2025, San Francisco, CA, USA, May 12-15, 2025*, pages 3460–3478. IEEE, 2025. doi:10.1109/SP61157.2025.00205.

A Related Work

We now briefly discuss further prior works in the field of model certification. A number of recent works aim to prove desirable model properties. In terms of *what* these works prove, they can be roughly categorized into proofs of training, inference, accuracy, and fairness. In terms of *how* the corresponding protocols work, recent works on certifiable ML can be categorized as follows:

Cryptographic approaches. A prolific line of research adapts various cryptographic techniques to certify properties such as accuracy, fairness, etc., without revealing the model's details. The most common technique is *zero-knowledge proofs* (zk proofs), which allow to formally prove that a model satisfies certain properties without revealing anything else about the model. They have been used to certify fairness [21, 42, 51, 57], inference [56], accuracy [56], and to prove that the model has been trained using a certain algorithm [2, 22, 38, 45] (without revealing the training data). Works such as Chang et al. [12] and Duddu et al. [18] use *secure multi-party computation* (MPC) [25, 52], a foundational cryptographic primitive introduced and refined through decades of research [5, 8, 13, 15, 16, 20, 23, 24, 27, 32, 34, 53]. MPC allows mutually distrusting parties to jointly compute on private inputs without revealing anything about the inputs apart from the outcome.

Black box auditing/Statistical testing. These approaches probe a model by submitting inputs, collecting outputs, and analyzing them for undesirable behavior. Works by Saleiro et al. [40] and Tramer et al. [47] use black-box testing to check for potential unfairness or bias, while Tan et al. [46] distill a new model to gain insight into the black box one.

Outside-the-box auditing. Here the model owner provides access to information beyond query responses, such as source code, documentation [36], hyperparameters, training data, deployment details, or internal evaluation results.

B Additional Preliminaries

B.1 Welch's t-test

Welch's t-test The goal of t-test is to determine whether the unknown population means of two groups are equal or not. That is, for random variables X and Y, it compares the following hypotheses on their means $\mu_X = \mathrm{E}[X]$ and $\mu_Y = \mathrm{E}[Y]$:

- Null Hypothesis H_0: $\mu_X = \mu_Y$
- Alternative Hypothesis H_1: $\mu_X \neq \mu_Y$

Assuming that X and Y independently follow Gaussian distributions with unknown variances, Welch's t-test proceeds as in Algorithm 4.

Algorithm 4 Welch's t-test

Input: $\mathcal{X} = \{x_i\}_{i\in[n]}$, $\mathcal{Y} = \{y_i\}_{i\in[m]}$, where $x_i \sim X$ and $y_i \sim Y$, and a significance level α **Output**: Null hypothesis H_0 (i.e., $\mu_X = \mu_Y$) or alternative hypothesis H_1 (i.e., $\mu_X \neq \mu_Y$)

1: Compute sampled means $\bar{x} = \frac{\sum_i x_i}{n}$ and $\bar{y} = \frac{\sum_i y_i}{m}$
2: Compute sampled variances $v_x = \frac{\sum_i (\bar{x}-x_i)^2}{n-1}$ and $v_y = \frac{\sum_i (\bar{y}-y_i)^2}{m-1}$.
3: Compute the test statistic $t = \frac{\bar{x}-\bar{y}}{\sqrt{v_x/n+v_y/m}}$
4: Compute the degree of freedom $d = \frac{(g_x+g_y)^2}{g_x^2/(n-1)+g_y^2/(m-1)}$, where $g_x = v_x/n$ and $g_y = v_y/m$
5: Obtain the critical value t_{cr} from the t-table, given d and α.
6: **If** $|t| < t_{cr}$ **return** H_0 **else return** H_1

B.2 Decision Trees

In our attack constructions we focus on decision tree models. Decision tree-based solutions are among the most popular machine learning algorithms, particularly known for their effectiveness in classification problems such as loan approval and fraud detection. A decision tree is trained by recursively partitioning the dataset from the root to the leaves. At each step, a split is determined by a splitting rule that aims to maximize an objective function, such as information gain. For prediction, the input follows a path from the root to a leaf, where at each internal node, the decision depends on whether the input satisfies the corresponding threshold (see Algorithm 5).

For completeness, in Algorithm 5 we present the algorithm for decision tree inference.

B.3 Security Properties of Zero-Knowledge Proofs

Let $\mathsf{ZKP} = (\mathcal{P}, \mathcal{V})$ be an interactive proof system for a relation $\mathcal{R} = \bigcup_{\lambda\in\mathbb{N}} \mathcal{R}_\lambda$. In what follows, we denote by PPT *probabilistic polynomial time*.

Completeness ZKP is (perfectly) *complete* if for any (x, w) satisfying $\mathcal{R}$, it holds that:

$$\Pr[1 \leftarrow \langle \mathcal{P}(\mathsf{w}), \mathcal{V}\rangle(\mathsf{x})] = 1.$$

Algorithm 5 Decision Tree Inference

Input: Decision tree h, input $\mathbf{a}$.
Output: Classification result.

```
Let cur := h.root                ▷ Set cur to be root of the tree
while cur is not a leaf do
    if a[cur.attr] < cur.thr then
        cur := cur.left. ▷ Set cur to be current node's left child
    else
        cur := cur.right.  ▷ Set cur to be current node's right child
return cur.class
```

Knowledge Soundness ZKP is (adaptively) *knowledge sound* with knowledge error κ if for any (stateful) PPT adversary $\mathcal{P}^* = (\mathcal{P}_0, \mathcal{P}_1)$, there exists an expected polynomial time extractor $\mathcal{E}$ such that the following holds:

$$p_{\text{ext}} \geq p_{\text{acc}} - \kappa$$

where

$$p_{\text{ext}} = \Pr\left[R_\lambda(x,w) = 1 : x \leftarrow \mathcal{P}_0(1^\lambda); w \leftarrow \mathcal{E}_{\mathcal{P}}(x)\right]$$
$$p_{\text{acc}} = \Pr\left[b = 1 : x \leftarrow \mathcal{P}_0(1^\lambda); b \leftarrow \langle \mathcal{P}_1, \mathcal{V}\rangle(x)\right]$$

where $\mathcal{E}$ has non-black-box access to $\mathcal{P}^*$. Informally, this means that any cheating prover must know a valid witness if it convinces verifier.

Zero-Knowledge Let $\text{view}_{\mathcal{V}}^{\mathcal{P}(w)}(x)$ be a string consisting of all the incoming messages that $\mathcal{V}$ receives from $\mathcal{P}$ during the interaction $\langle \mathcal{P}(w), \mathcal{V}\rangle(x)$, and $\mathcal{V}$'s random coins. Π is (honest verifier) *zero-knowledge* if there exists a PPT simulator $\mathcal{S}$ such that for any adversary $\mathcal{A}$ and any $(x,w) \in \mathcal{R}_\lambda$, the following is negligible in λ.

$$\left| \begin{array}{l} \Pr\left[b = 1 : b \leftarrow \mathcal{A}(\text{view}_{\mathcal{V}}^{\mathcal{P}(w)}(x)) \right] \\ -\Pr\left[b = 1 : \text{view}' \leftarrow \mathcal{S}(x); \quad b \leftarrow \mathcal{A}(\text{view}')\right] \end{array} \right|$$

Informally, this means that the protocol execution reveals no information to $\mathcal{V}$ about w.

B.4 Security Properties of Commitment Schemes

Let Commit be a commitment scheme. For simplicity, we omit the key generation algorithm Gen and present a class of the simplest commitments whose openings are checked by recomputing and comparing (e.g., hash commitment $H(m||\rho)$). More generally, some commitment schemes require a separate verification algorithm Verify to check the validity of a commitment given some *decommitment* information. Our auditing framework can be extended to such schemes by having the model provider prove the knowledge of the decommitment information in zero knowledge.

Definition 5 (Commitment Scheme)**.** *A commitment scheme is an algorithm* Commit*, which is executed as* $\text{com} \leftarrow \text{Commit}(m;\rho)$*. It takes as input a message* $m \in \{0,1\}^{\ell_m(\lambda)}$*, a uniformly sampled randomness* $\rho \in \{0,1\}^{\ell_r(\lambda)}$*, and returns a commitment* $\text{com} \in \{0,1\}^{\ell_c(\lambda)}$*. Here* ℓ_m, ℓ_r, ℓ_c *are some polynomials in* λ*, the security parameter (determining the desired level of security).*

Binding Commit is *computationally binding* if for any PPT adversary $\mathcal{A}$, the following is negligible in λ:

$$\Pr\left[\begin{array}{l} \text{com} = \text{Commit}(m;\rho) \\ \text{com} = \text{Commit}(m';\rho') : (\text{com}, m, m', \rho, \rho') \leftarrow \mathcal{A}(1^\lambda) \\ \quad m \neq m' \end{array}\right]$$

Hiding Commit is *computationally hiding* if for any PPT adversary $\mathcal{A}$, the following is negligible in λ:

$$\left| \Pr\left[b = b' : \begin{array}{r} (m_0, m_1) \leftarrow \mathcal{A}(1^\lambda); \\ \rho \leftarrow \{0,1\}^{\ell_r(\lambda)}; b \leftarrow \{0,1\}; \\ \text{com} \leftarrow \text{Commit}(m_b;\rho); b' \leftarrow \mathcal{A}(\text{com}) \end{array}\right] - \frac{1}{2} \right|$$

C Proof of Theorem 1

Proof. First, let us show that if points a and b, where $a_i < c_i < b_i$ and $c_j = b_j$ for all $j \neq i$, get sorted into the same decision tree leaf, then c is also sorted into that same leaf. Notice that in order for two points x and y to get sorted into different leaves, there must be a node which splits on a feature i such that $x_i \neq y_i$ and $x_i \leq t \leq y_i$ or $y_i \leq t \leq x_i$ where t is the threshold to split upon. Then if b and c were sorted into separate leaves, there must be a node on the path that b takes through the tree that splits on feature i with a threshold t that satisfies $c_i \leq t \leq b_i$. However, such a node would also sort a distinctly from b, so such a node cannot occur. Thus, b and c must be sorted into the same leaf.

Now, note that as we train the tree until its leafs are homogeneous, every datapoint in S'_{train} must be classified correctly (according to the label we assigned to it in Algorithm 1). Further, since the only datapoints in S'_{train} with non-zero labels are datapoints from S_{audit}, for every leaf in $\mathcal{T}$ that is associated with a non-zero class, we have at least one $r \in S_{audit}$ that gets sorted into this leaf.

Consider $x \in \mathbb{R}^d$ such that $||x - r||_\infty \geq \varepsilon$ for all $r \in S_{audit}$. Say $\mathcal{T}(x) = 1$, i.e., there exists a leaf such that x belongs to this leaf and the leaf corresponds to class one. Consider $r \in S_{audit}$ that belongs to this leaf (by above, such r exists). By the definition of the L-infinity norm there exists some dimension i where $|r_i - x_i| > \varepsilon$. Suppose $r_i - x_i > \varepsilon$. Notice that there is a

point $r-\varepsilon\vec{b}_i \in S'_{train}$ which satisfies that $(r-\varepsilon\vec{b}_i)_j = r_j$ for all $j \neq i$, and where $x_i < (r-\varepsilon\vec{b}_i)_i < r_i$. Then by above, $r-\varepsilon\vec{b}_i$ must be sorted into the same leaf as r and x. But $r-\varepsilon\vec{b}_i$ has label $g(r)=0$, while for x holds $\mathcal{T}(x)=1$. Thus, we found a contradition. The same argument holds if $x_i - r_i > \varepsilon$, but using the point $r+\varepsilon\vec{b}_i$ instead of $r-\varepsilon\vec{b}_i$. □

D Attack Detection

While proof of training alone cannot detect the attack above (as it relies on training the decision tree entirely honestly), nor can a black-box audit where the model owner knows the audit data before training time, we might still hope to detect when these attacks occur. For example, we might hope to conduct statistical tests on the training data to determine if it was honestly sampled from the underlying distribution or if it was adversarially constructed. In such a case, we cannot directly compare the training data to the true distribution of real data because the underlying distribution is not fully known to the auditor. Instead, we must compare the training data with a sample from that distribution. In the most simple case, this sample is the reference set S_{audit}.

We argue that under a certain family of functions, our constructed training set is indistinguishable from S_{audit}.

Definition 6. *Suppose $\vec{\alpha}$ is a set of bins over d dimensions. Then $H_{\vec{\alpha}} : (\mathbb{R}^d \times \{0,1\})^* \to \mathcal{H}$ is the function which takes databases over d features and a binary classification to their normalized histogram with bins $\vec{\alpha}$.*

Definition 7. *A function $f : (\mathbb{R}^d \times \{0,1\})^* \to \mathbb{R}$ is called (γ, c)-magnitude insensitive if there exists a choice of bins $\vec{\alpha}$ and function $f' : \mathcal{H} \to \mathbb{R}$ such that $|f(D) - f'(H_{\vec{\alpha}}(D))| < \gamma$ for all $D \in (\mathbb{R}^d \times \{0,1\})^*$ and $|f'(H_{\vec{\alpha}}(D)) - f'(H_{\vec{\alpha}}(D||r))| \leq \frac{c}{|D|}$ for all $D \in (\mathbb{R}^d \times \{0,1\})^*$ and $r \in \mathbb{R}^d \times \{0,1\}$.*

Theorem 2. *If f is (γ, c)-magnitude insensitive, then $|f(S_{audit}) - f\left(S^k_{audit}||\delta\right)| \leq \varepsilon$ for any $\varepsilon > 2\gamma$ and $k \geq \frac{2dc}{\varepsilon - 2\gamma}$, where δ is the additional training data created by Algorithm 1 when run with input $S_{audit}, d, \varepsilon, g$ for any g.*

Proof. We will write f' to be the γ-approximation of f guaranteed to exist by the fact that f is (γ, c)-magnitude insensitive. Observe that because $H_{\vec{\alpha}}$ takes databases to their normalized histograms, $H_{\vec{\alpha}}(S_{audit}) = H_{\vec{\alpha}}\left(S^k_{audit}\right)$, because the non-normalized histograms of the two databases are simply scaled versions of one another.

Next, it will be helpful to show that for any two databases $D_1, D_2 \in (\mathbb{R}^d \times \{0,1\})^*$, we have $|f'(H_{\vec{\alpha}}(D_1)) - f'(H_{\vec{\alpha}}(D_1||D_2))| \leq c\frac{|D_2|}{|D_1|}$. Let us write $D_2 = d_1||d_2||\ldots||d_{|D_2|}$. Then we get that

$$
\begin{aligned}
&|f'(H_{\vec{\alpha}}(D_1)) - f'(H_{\vec{\alpha}}(D_1||D_2))| \\
&= |f'(H_{\vec{\alpha}}(D_1)) - f'(H_{\vec{\alpha}}(D_1||d_1)) + f'(H_{\vec{\alpha}}(D_1||d_1)) - \ldots \\
&\quad + f'(H_{\vec{\alpha}}(D_1||d_1||d_2||\ldots||d_{|D_2|-1})) - f'(H_{\vec{\alpha}}(D_1||D_2))| \\
&\leq |f'(H_{\vec{\alpha}}(D_1)) - f'(H_{\vec{\alpha}}(D_1||d_1))| + \\
&\quad |f'(H_{\vec{\alpha}}(D_1||d_1)) - f'(H_{\vec{\alpha}}(D_1||d_1||d_2)| + \ldots \\
&\quad + |f'(H_{\vec{\alpha}}(D_1||d_1||d_2||\ldots||d_{|D_2|-1})) - f'(H_{\vec{\alpha}}(D_1||D_2))| \\
&\leq \frac{c}{|D_1|} + \frac{c}{|D_1|+1} + \ldots + \frac{c}{|D_1|+|D_2|-1} \\
&\leq c\frac{|D_2|}{|D_1|}
\end{aligned}
$$

Then we can apply this to S^k_{audit} and $S^k_{audit}||\delta$; recall that $|\delta| = 2d|S_{audit}|$. Then we see that

$$
\begin{aligned}
&\left|f'\left(H_{\vec{\alpha}}\left(S_{audit}\right)\right) - f'\left(H_{\vec{\alpha}}\left(S^k_{audit}||\delta\right)\right)\right| \\
&= \left|f'\left(H_{\vec{\alpha}}\left(S^k_{audit}\right)\right) - f'\left(H_{\vec{\alpha}}\left(S^k_{audit}||\delta\right)\right)\right| \\
&\leq c\frac{2d|S_{audit}|}{k|S_{audit}|} \\
&\leq c\frac{2d}{\left(\frac{2dc}{\varepsilon-2\gamma}\right)} = \varepsilon - 2\gamma
\end{aligned}
$$

We have two cases now.

Case 1: $f'\left(H_{\vec{\alpha}}\left(S_{audit}\right)\right) \geq f'\left(H_{\vec{\alpha}}\left(S^k_{audit}||\delta\right)\right)$. Then we have

$$
\begin{aligned}
\varepsilon - 2\gamma &\geq f'\left(H_{\vec{\alpha}}\left(S_{audit}\right)\right) - f'\left(H_{\vec{\alpha}}\left(S^k_{audit}||\delta\right)\right) \\
&= f(S_{audit}) - f(S_{audit}) + f'(H_{\vec{\alpha}}(S_{audit})) \\
&\quad - f(S^k_{audit}||\delta) + f(S^k_{audit}||\delta) - f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) \\
&\geq f(S_{audit}) - |f(S_{audit}) - f'(H_{\vec{\alpha}}(S_{audit}))| \\
&\quad - f(S^k_{audit}||\delta) - |f(S^k_{audit}||\delta) - f'(H_{\vec{\alpha}}(S^k_{audit}||\delta))| \\
&\geq f(S_{audit}) - \gamma - f(S^k_{audit}||\delta) - \gamma
\end{aligned}
$$

and so we see that $\varepsilon \geq f(S_{audit}) - f(S^k_{audit}||\delta)$. We also have

$$
\begin{aligned}
&f(S_{audit}) - f(S^k_{audit}||\delta) \\
&= f'(H_{\vec{\alpha}}(S_{audit})) - f'(H_{\vec{\alpha}}(S_{audit})) + f(S_{audit}) \\
&\quad - f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) + f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) - f(S^k_{audit}||\delta) \\
&\geq f'(H_{\vec{\alpha}}(S_{audit})) - |f'(H_{\vec{\alpha}}(S_{audit})) - f(S_{audit})| \\
&\quad - f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) - |f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) - f(S^k_{audit}||\delta)| \\
&\geq f'(H_{\vec{\alpha}}(S_{audit})) - \gamma - f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) - \gamma \\
&\geq -2\gamma \\
&> -\varepsilon
\end{aligned}
$$

Then $|f(S_{audit}) - f(S^k_{audit}||\delta)| \leq \varepsilon$.

Case 2: $f'(H_{\vec{\alpha}}(S_{audit})) \leq f'(H_{\vec{\alpha}}(S^k_{audit}||\delta))$. Then we have

$$\begin{aligned}
\varepsilon - 2\gamma &\geq f'\left(H_{\vec{\alpha}}\left(S^k_{audit}||\delta\right)\right) - f'\left(H_{\vec{\alpha}}\left(S_{audit}\right)\right) \\
&= f(S^k_{audit}||\delta) - f(S^k_{audit}||\delta) + f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) \\
&\quad - f(S_{audit}) + f(S_{audit}) - f'(H_{\vec{\alpha}}(S_{audit})) \\
&\geq f(S^k_{audit}||\delta) - |f(S^k_{audit}||\delta) - f'(H_{\vec{\alpha}}(S^k_{audit}||\delta))| \\
&\quad - f(S_{audit}) - |f(S_{audit}) - f'(H_{\vec{\alpha}}(S_{audit}))| \\
&\geq f(S^k_{audit}||\delta) - \gamma - f(S_{audit}) - \gamma
\end{aligned}$$

and so we see that $\varepsilon \geq f(S^k_{audit}||\delta) - f(S_{audit})$. We also have

$$\begin{aligned}
& f(S^k_{audit}||\delta) - f(S_{audit}) \\
&= f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) - f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) + f(S^k_{audit}||\delta) \\
&\quad - f'(H_{\vec{\alpha}}(S_{audit})) + f'(H_{\vec{\alpha}}(S_{audit})) - f(S_{audit}) \\
&\geq f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) - |f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) + f(S^k_{audit}||\delta)| \\
&\quad - f'(H_{\vec{\alpha}}(S_{audit})) - |f'(H_{\vec{\alpha}}(S_{audit})) - f(S_{audit})| \\
&\geq f'(H_{\vec{\alpha}}(S^k_{audit}||\delta)) - \gamma - f'(H_{\vec{\alpha}}(S_{audit})) - \gamma \\
&\geq -2\gamma \\
&\geq -\varepsilon
\end{aligned}$$

Then $|f(S_{audit}) - f(S^k_{audit}||\delta)| \leq \varepsilon$. □

This theorem does not suggest that it is completely impossible to detect the attack given in Algorithm 1. Rather, it only precludes detection by a certain class of functions. However, we argue that this class is expansive and covers many intuitive approaches.

The sole requirement for the audit metric f is that it must be approximable by f' which satisfies three properties. Firstly, f' operates over histograms for some choice of bins $\vec{\alpha}$. This is a necessary condition, as if f were not approximable by a function over a binning of the training data, we could drastically change the audit outcome by simply adding a small amount of noise to the data. Next, f' must be relatively insensitive to additional data. The intuition here is that no individual datapoint should dramatically change the outcome of the audit. Finally, f' operates over normalized histograms. This property is necessary for the proof to go through, but is satisfied by many intuitive audit metrics. For example, the mean and standard deviation of a feature (even conditioned on any arbitrary set of features) are approximable from a normalized histogram.

Lemma 1. *Let $\mu_j(D)$ be the mean of (bounded) feature j of a dataset D. Then for every $\gamma > 0$, $\mu_j(D)$ is $(\gamma, M-m)$-magnitude insensitive, where B is the set of bins in the histogram and M, m are an upper and lower bound on possible j-values respectively.*

Proof. Notice that $\mu_j(D) \approx \sum_{i \in B} p_i x_{j,i}$ where B is the set of bins in the histogram, p_i is the height of bin i in the normalized histogram of D, and $x_{j,i}$ is the j-value of bin i. Let us show that for any $\gamma > 0$, there exists a binning of the data such that this is a γ-approximation of $\mu_j(D)$. Let the bins in feature j have width γ. Then for each datapoint d with j value j_d, bin i, and binned j-value $x_{j,i}$, we have that $|x_{j,i} - j_d| \leq \gamma$. Then

$$\begin{aligned}
\sum_{i \in B} p_i x_{j,i} &= \sum_{i \in B} \frac{c_i}{|D|} x_{j,i} \\
&= \sum_{d \in D} \frac{1}{|D|} x_{j,i} \\
\implies \left| \sum_{i \in B} p_i x_{j,i} - \sum_{d \in D} \frac{1}{|D|} j_d \right| &= \left| \sum_{d \in D} \frac{1}{|D|} x_{j,i} - \sum_{d \in D} \frac{1}{|D|} j_d \right| \\
&= \left| \frac{1}{|D|} \sum_{d \in D} (x_{j,i} - j_d) \right| \\
&\leq \frac{1}{|D|} \sum_{d \in D} |x_{j,i} - j_d| \\
&\leq \frac{1}{|D|} \sum_{d \in D} \gamma \\
&= \gamma
\end{aligned}$$

Next, let us show that the sensitivity of our approximation of μ_j is upper bounded by $\frac{M-m}{|D|}$. Notice that by adding a single point, one histogram bin will increase by 1 and the rest will be unchanged. Then for every bin k,

$$\begin{aligned}
&\sum_{i \in B} \frac{c_i}{|D|+1} x_{j,i} + \frac{1}{|D|+1} x_{j,k} - \sum_{i \in B} \frac{c_i}{|D|} x_{j,i} \\
&\quad = \sum_{i \in B} c_i x_{j,i} \left(\frac{1}{|D|+1} - \frac{1}{|D|} \right) + \frac{x_{j,k}}{|D|+1} \\
&\quad = -\left(\sum_{i \in B} \frac{c_i x_{j,i}}{|D|^2 + |D|} \right) + \frac{x_{j,k}}{|D|+1} \\
&\quad \leq -\left(\frac{m}{|D|+1} \right) + \frac{M}{|D|+1} \\
&\quad \leq \frac{M-m}{|D|} \\
&\sum_{i \in B} \frac{c_j}{|D|+1} x_{j,i} + \frac{1}{|D|+1} x_{j,k} - \sum_{i \in B} \frac{c_j}{|D|} x_{j,i} \\
&\quad = -\left(\sum_{i \in B} \frac{c_i x_{j,i}}{|D|^2 + |D|} \right) + \frac{x_{j,k}}{|D|+1} \\
&\quad \geq -\left(\frac{M}{|D|+1} \right) + \frac{m}{|D|+1} \\
&\quad \geq \frac{m-M}{|D|}
\end{aligned}$$

So we have that the sensitivity is no greater than $\frac{M-m}{|D|}$. □

We will proceed to use this fact to show that Welch's t-test will fail to detect this attack.

Corollary 2. *(Restatement of Corollary 1) Given an audit dataset S_{audit} and significance level α, we can use Algorithm 1 to construct a training dataset S'_{train} such that for any feature j, S'_{train} passes Welch's t-test when its values in feature j are compared to those of S_{audit} with significance level α.*

Before we can prove this corollary, we will need a lemma which bounds the concentration of the Student's t-distribution.

Lemma 2. *If X and Z are random variables drawn independently from the Student's t-distribution with ν degrees of freedom and the standard normal distribution respectively, then for every $t > 0$, we have*

$$\Pr[|X| < t] \leq \Pr[|Z| < t]$$

Proof. We will write $F_X(t)$ to denote the CDF of random variable X evaluated at t, and $f_X(t)$ the PDF. We will also write $\mathbb{E}_X(g(X))$ to be the expected value of $g(X)$ with randomness over X. Let us begin by demonstrating that for all $t < 0$, we have $F_X(t) > F_Z(t)$. First, recall that if W and Y are drawn from the χ^2 distribution with ν degrees of freedom and the standard normal distribution respectively, then $Y\sqrt{\frac{\nu}{W}}$ is distributed according to the Student's t-distribution with ν degrees of freedom, so let us write $X = Y\sqrt{\frac{\nu}{W}}$. Then according to the law of total probability, we have

$$\begin{aligned} F_X(t) &= \int_0^\infty F_Y\left(t\sqrt{\frac{w}{\nu}}\right) f(w)dw \\ &= \mathbb{E}_W\left(F_Y\left(t\sqrt{\frac{W}{\nu}}\right)\right) \end{aligned}$$

Notice that $\frac{d^2}{dt^2}F_Y(t) = \frac{d}{dt}f_Y(t) = \frac{d}{dt}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} = -\frac{t}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} > 0$ when $t < 0$. Then since $t\sqrt{\frac{W}{\nu}}$ must be less than 0, we can apply Jensen's inequality to get

$$\begin{aligned} F_X(t) &= \mathbb{E}_W\left(F_Y\left(t\sqrt{\frac{W}{\nu}}\right)\right) \\ &\geq F_Y\left(\mathbb{E}_W\left(t\sqrt{\frac{W}{\nu}}\right)\right) \\ &= F_Y\left(t\mathbb{E}_W\left(\sqrt{\frac{W}{\nu}}\right)\right) \end{aligned}$$

Then since $\frac{d^2}{du^2}\sqrt{u} = -\frac{1}{4\sqrt{u^3}} \leq 0$, we get that $\mathbb{E}_W\left(\sqrt{\frac{W}{\nu}}\right) \leq \sqrt{\frac{\mathbb{E}_W(W)}{\nu}} = \sqrt{\frac{\nu}{\nu}} = 1$. So because $t < 0$, we can see that $t\mathbb{E}_W\left(\sqrt{\frac{W}{\nu}}\right) \geq t$, and since $F_Y(u)$ is increasing, we get

$$\begin{aligned} F_X(t) &\geq F_Y\left(t\mathbb{E}_W\left(\sqrt{\frac{W}{\nu}}\right)\right) \\ &\geq F_Y(t) \end{aligned}$$

Since f_X and f_Y are both symmetric about $t = 0$, it then follows by a symmetric argument that for all $t > 0$, $F_X(t) \leq F_Y(t)$. Then we see that for any $t > 0$,

$$\begin{aligned} \Pr[|X| < t] &= F_X(t) - F_X(-t) \\ &\leq F_Y(t) - F_Y(-t) \\ &= \Pr[|Y| < t] \\ &= \Pr[|Z| < t] \end{aligned}$$

Because Y and Z are independently and identically distributed. □

We are now ready to prove Corollary 2.

Proof of Corollary 2. A pair of datasets D_1, D_2 pass Welch's t-test on feature j if

$$\frac{|\mu_j(D_1) - \mu_j(D_2)|}{\sqrt{\frac{\sigma_1^2}{|D_1|} + \frac{\sigma_2^2}{|D_2|}}} \leq T_{\alpha,\nu}$$

where α is the desired significance level, ν is the degrees of freedom in the datasets, and $T_{\alpha,\nu}$ is the unique value such that

$$\Pr_{x\sim t(\nu)}[|x| \geq T_{\alpha,\nu}] = \alpha$$

where $t(\nu)$ is the Student's t-distribution with ν degrees of freedom. In our case, the t-test compares the reference dataset S_{audit} with the training dataset S'_{train}.

The value of ν, and thus the value of $T_{\alpha,\nu}$, depends on the size of the datasets, with the threshold $T_{\alpha,\nu}$ decreasing as the datasets grow large. However, we will use Lemma 2 to give a lower bound for $T_{\alpha,\nu}$ which is constant with respect to $|S'_{train}|$. Then, we will show that by Lemma 1 and Theorem 2 we can use Algorithm 1 to construct a malicious training dataset S'_{train} which maintains an arbitrarily small test statistic, and in particular, a dataset such that the test statistic is below the lower bound on the threshold.

First, let us establish a lower bound on $T_{\alpha,\nu}$. Let us define T'_α to be the unique positive value such that

$$\Pr_{Z\sim\mathcal{N}(0,1)}[|Z| \geq T'_\alpha] = \alpha$$

Then recall that Lemma 2 gives us that

$$\Pr_{X\sim t(\nu)}[|X| < T'_\alpha] \leq \Pr_{Z\sim\mathcal{N}(0,1)}[|Z| < T'_\alpha]$$

If we write f_X and f_Z to represent the probability density functions (PDFs) of X and Z respectively, then we get equivalently that

$$\int_{-T'_\alpha}^{T'_\alpha} f_X(u)du \leq \int_{-T'_\alpha}^{T'_\alpha} f_Z(u)du$$

Then we see that

$$\begin{aligned}
\Pr_{Z\sim\mathcal{N}(0,1)}[|Z|\geq T'_\alpha] &= \Pr_{X\sim t(\nu)}[|X|\geq T_{\alpha,\nu}] \\
\implies \int_{-T'_\alpha}^{T'_\alpha} f_Z(u)du &= \int_{-T_{\alpha,\nu}}^{T_{\alpha,\nu}} f_X(u)du \\
&= \int_{-T_{\alpha,\nu}}^{-T'_\alpha} f_X(u)du + \int_{-T'_\alpha}^{T'_\alpha} f_X(u)du \\
&\quad + \int_{T'_\alpha}^{T_{\alpha,\nu}} f_X(u)du \\
&\leq \int_{-T_{\alpha,\nu}}^{-T'_\alpha} f_X(u)du + \int_{-T'_\alpha}^{T'_\alpha} f_Z(u)du \\
&\quad + \int_{T'_\alpha}^{T_{\alpha,\nu}} f_X(u)du \\
\implies 0 &\leq \int_{-T_{\alpha,\nu}}^{-T'_\alpha} f_X(u)du + \int_{T'_\alpha}^{T_{\alpha,\nu}} f_X(u)du
\end{aligned}$$

Then because $f_X(x)$ is symmetric about $x=0$, this yields

$$2\int_{T'_\alpha}^{T_{\alpha,\nu}} f_X(u)du \geq 0$$

and thus

$$\int_{T'_\alpha}^{T_{\alpha,\nu}} f_X(u)du \geq 0$$

Now recall the simple result from calculus that states that if g is positive valued, then

$$\int_a^b g(x)dx \geq 0 \iff a \leq b$$

Then because f_X is positive-valued, our prior result entails that $T_{\alpha,\nu} \geq T'_\alpha$, so T'_α is a lower bound on $T_{\alpha,\nu}$ that does not depend on $|S'_{train}|$.

Next, observe that the test statistic for Welch's t-test has the following upper bound:

$$\frac{|\mu_j(S'_{train}) - \mu_j(S_{audit})|}{\sqrt{\frac{\sigma^2_{train}}{|S'_{train}|} + \frac{\sigma^2_{audit}}{|S_{audit}|}}} \leq \frac{|\mu_j(S'_{train}) - \mu_j(S_{audit})|}{\sqrt{\frac{\sigma^2_{audit}}{|S_{audit}|}}}$$

Furthermore, Lemma 1 implies that for any $\varepsilon > 0$, we can choose $\gamma < \frac{\varepsilon}{2}$ such that μ_j is (γ, c)-magnitude insensitive, and so by Theorem 2, Algorithm 1 yields a dataset S'_{train} such that $|\mu_j(S'_{train}) - \mu_j(S_{audit})| \leq \varepsilon$ when appropriately parameterized. Then let $\varepsilon = T'_\alpha \frac{\sigma_{audit}}{2\sqrt{|S_{audit}|}}$. This produces the result that

$$\begin{aligned}
\frac{|\mu_j(S'_{train}) - \mu_j(S_{audit})|}{\sqrt{\frac{\sigma^2_{train}}{|S'_{train}|} + \frac{\sigma^2_{audit}}{|S_{audit}|}}} &\leq \frac{2\varepsilon}{\sqrt{\frac{\sigma^2_{audit}}{|S_{audit}|}}} \\
&= \frac{2}{\sqrt{\frac{\sigma^2_{audit}}{|S_{audit}|}}} T'_\alpha \frac{\sigma_{audit}}{2\sqrt{|S_{audit}|}} \\
&= T'_\alpha \\
&\leq T_{\alpha,\nu}
\end{aligned}$$

which passes the t-test for feature j. Finally, by choosing $k = \max_j \frac{4d(M_j - m_j)\sqrt{|S_{audit}|}}{T'_\alpha \sigma_{audit,j}}$ we get for every feature i that $|\mu_i(S'_{train}) - \mu_i(S_{audit})| \leq 2\min_j T'_\alpha \frac{\sigma_{audit,j}}{2\sqrt{|S_{audit}|}} \leq 2T'_\alpha \frac{\sigma_{audit,i}}{2\sqrt{|S_{audit}|}}$, so S'_{train} passes the t-test for feature i. □

E Case Study

We now discuss a number of state of the art works that consider the problem of privacy-preserving auditing. These works are focused on different auditing functions (accuracy, fairness, etc), different types of machine learning models, and their security models they use are not necessarily aligned. We now briefly outline the techniques and security guarantees that are claimed in each of the works. Our goal is not to provide an exhaustive survey, but rather to illustrate the landscape through recent works that are broadly representative of the field—even though they span different years, venues, and communities (ranging from machine learning to security).

E.1 zkCNN: Zero Knowledge Proofs for Convolutional Neural Network Predictions and Accuracy

Goal and Solution Details. [35] propose zkCNN, a zero-knowledge proof protocol for inference and accuracy of convolutional neural networks (CNNs). The core contribution is a novel sumcheck protocol (which is the key ingredient in many zero-knowledge system) that is tailored to two-dimensional convolutions.

Security Model. zkCNN considers the standard setting with a prover and a verifier. Either party can be malicious. [35]'s security definition for inference is a zero-knowledge-style definition, and the scheme is required to satisfy correctness, soundness, and zero-knowledge. Similar to [56], [35]'s soundness intuitively states that a prover should not be able to output a commitment to a model and provide a proof π, prediction y and datapoint X such that the verifier accepts the proof, and at the same time, the committed model's prediction for X is not equal to y. If instantiated with a specific commitment scheme, [35]'s scheme further satisfies knowledge soundness, the stronger version of soundness where there exists an extractor to extract the CNN parameters from a valid proof and prediction with overwhelming probability. [35] do not provide a security definition for their proof of accuracy.

Discussion. As [35] do not give a security definition for their proof of accuracy, the formal security guarantee they provide is not fully clear. However, the authors indicate that their scheme can be used to prove the accuracy on a public dataset. This scenario falls within our framework of definition 3, and is vulnerable to the same style of attack as outlined in §4.1.

Expanding/Restoring Security. Similar to Zhang et al. [56], [35]'s construction can be adapted using our techniques from

§ 5.3 to satisfy the guarantee with respect to new datasets drawn from the same distribution. In particular, Liu et al.'s scheme can be plugged in as the ZKP of our protocol Π_{csp} in §5.2.

E.2 Confidential-PROFITT: Confidential PROof of FaIr Training of Trees

Goal and Solution Details. Shamsabadi et al. [42] propose Confidential-PROFITT, a framework for certifying fairness of decision trees while preserving confidentiality of both the model and the training data. Confidential-PROFITT consists of a zero-knowledge-friendly decision tree learning algorithm that, when executed honestly, enforces fairness by design—up to a tunable degree controlled by a parameter. On top of this, Confidential-PROFITT designs a zero-knowledge proof system to verify fairness of a decision tree. The proof requires the model provider to commit to both the model and its training data, then prove in zero-knowledge that the paths taken by the committed training points through the (committed) decision tree satisfy specified fairness bounds. In terms of fairness metrics, Confidential-PROFITT supports demographic parity and equalized odds.

Security Model. Confidential-PROFITT considers a malicious model provider (that, however, is assumed to commit to the training data honestly) and a malicious auditor, and obtains standard zero-knowledge proof properties (correctness, soundness, zero-knowledge) with respect to a statement that can be summarized roughly as follows "With respect to a private dataset *chosen by the model provider*, the committed model satisfies certain fairness guarantees".

Discussion. Confidential-PROFITT assumes that the model provider honestly commits to the training data. Under this assumption, the corresponding zero-knowledge proof certifies that the resulting model inherits the fairness guarantees of the fair learning algorithm introduced in Confidential-PROFITT (which the authors show indeed improves fairness). However, if the provider is not restricted to committing to the true training data, Confidential-PROFITT is vulnerable to data-forging attacks, as the provider can choose the audit dataset before committing to the model.

Expanding/Restoring Security. We believe Confidential-PROFITT's zero-knowledge proof of fairness could potentially be combined with our techniques from § 5.3 if one would swap the training data for the data sampled by the auditor. However, the adaptation is not as straight-forward as that for Zhang et al. [56] and Liu et al. [35]. This is because Confidential-PROFITT verifies not only the actual fairness guarantee, but also additional fairness checks for intermediate nodes in the decision tree. A careful analysis is required to ensure that the generalization of fairness properties (similar to what we show for demographic parity in § F.2.2)also extends to intermediate nodes. We leave this as an interesting direction for future work.

E.3 P2NIA: Privacy-Preserving Non-Iterative Auditing

Goal and Solution Details. Bourrée et al. [9] propose a novel auditing scheme that enables one-shot verification of a model's group fairness while preserving privacy for both parties: the model provider is not required to open-source the model, and the auditor need not disclose any private information to support the audit. The main contribution of [9] is a mechanism that enables auditing without requiring the auditor to supply the audit dataset. Specifically, the model provider supplies a dataset together with the corresponding predictions (both in the clear), which the auditor then uses to verify the fairness condition. To construct this dataset, model provider draws on a portion of its internal training data. To preserve confidentiality of this data, it is not shared directly. Instead, model provider feeds it into a synthetic data generation algorithm, and the resulting synthetic dataset is what is sent to the auditor.

Security Model. The work does not provide a formal security model. It is set up in the black-box setting and assumes that the auditor does not know the distribution of the model owner's training data.

Discussion. As [9] do not utilize cryptographic techniques to prove that the outputs actually correspond to the given inputs, the prover can easily cheat by simply adjusting the labels it supplies for the constructed dataset. However, even if one were to strengthen the scheme by adding a secure proof of training (e.g., [38]) together with inference proofs (as in [56]), the fact that the model owner knows the dataset that is being used for the audit means that the solution falls within our framework of Definition 3, and is thus vulnerable to data-forging attacks. An interesting open question would be to see if, since in this scenario the model owner not only knows, but directly influences the audit dataset, there can be an even simpler attack.

E.4 OATH: Efficient and Flexible Zero-Knowledge Proofs of End-to-End ML Fairness

Goal and Solution Details. [21] present OATH, a model-agnostic fairness auditing framework. The core idea in OATH is to leverage clients (who query the model during deployment) to participate in the auditing process. OATH operates in two phases: (i) a certification protocol between the model provider and the auditor, and (ii) a query authentication protocol involving model provider, inference clients, and auditor (dubbed verifier in OATH). The first phase follows the standard certification flow we describe in §2. In the second phase, the auditor receives commitments to client queries and the corresponding model predictions. These commitments can later be verified in zero knowledge for fairness, correctness, and consistency with the certified model.

Security Model. OATH considers three fully malicious entities: a model provider, inference clients, and an auditor. These parties are assumed not to collude with each other. The auditor assesses model fairness both with respect to the calibration dataset and the clients queries. The system provides standard correctness, soundness, and zero-knowledge with respect to these two datasets.

Discussion. The calibration dataset which is used in the certification protocol between the model provider and the auditor might be supplied by either party. If the calibration dataset is chosen by the prover, same as P2NIA and Confidential-PROFITT, the corresponding fairness check is vulnerable to data forging. However, in contrast to prior works, OATH can fall back on guarantees based on client's queries.

E.5 FairProof: Confidential and Certifiable Fairness for Neural Networks

Goal and Solution Details. [51] propose FairProof, a fairness certification approach that maintains confidentiality of the model. In contrast to Confidential-PROFITT and OATH, which focus on group fairness metrics, FairProof considers local individual fairness. This allows [51] to issue a personalized certificate to every client.

Security Model. FairProof system involves a malicious model provider and malicious clients (who wish to learn model details/training data), and considers standard correctness, soundness, and zero-knowledge properties. The corresponding statement is roughly as follows: "Given a datapoint x, the model's output is y and a lower bound on an individual fairness parameter for x is ε_x".

Discussion. The usage of a specific fairness metric (local individual fairness) allows FairProof to provide per-client certificates of fairness, and escape the problems that arise from the usage of reference datasets (including vulnerability to data-forging attacks). On the flip side, FairProof requires to generate fairness certificates during deployment and does not provide any fairness guarantees prior to deployment.

E.6 Scaling up Trustless DNN Inference with Zero-Knowledge Proofs

Goal and Solution Details. Kang et al. [33] propose a zero-knowledge-based framework for verifying DNN inference and accuracy. Their key contribution is a careful translation of DNN specifications into arithmetic circuits suitable for zero-knowledge proofs. The system also introduces economic incentives to support ML-as-a-service. Concretely, when verifying accuracy, the model provider first commits to the model, and the client commits to the test set. Both parties then deposit monetary collateral into an escrow. The client reveals the test set, and the provider must produce a zero-knowledge proof that the committed model meets the claimed accuracy. If the provider fails or refuses to prove the required accuracy, it forfeits its collateral; otherwise, the client pays for the service.

Security Model. Kang et al. study the standard two-party setting with a *prover* (model provider) and a *verifier* (client), either of whom may be malicious. Cryptographically, they aim for the standard zero-knowledge proof properties: *completeness*, *knowledge soundness*, and *zero knowledge*. They further consider incentives, showing that—under certain assumptions—honest model providers and clients are motivated to participate in the accuracy verification protocol, while malicious parties are discouraged.

Discussion. In terms of cryptographic guarantees, Kang et al. gets the core design right: their protocol for proofs of accuracy closely follows the framework outlined in §5 and is not vulnerable to our data-forging attacks. However, [33] provide no formal guarantees about accuracy on data outside the audited set.It would be interesting to perform an analysis similar to that in §F.2.1 given their constraints.

E.7 ezDPS: An Efficient and Zero-Knowledge Machine Learning Inference Pipeline

Goal and Solution Details. Wang and Hoang [48] introduce ezDPS, a pipeline for zero-knowledge proofs of inference correctness and accuracy above a specified threshold. They construct arithmetic circuit gadgets for key ML operations, including exponentiation, absolute value, and array max/min, and further devise optimized methods for proving Discrete Wavelet Transform, Principal Component Analysis, and multi-class Support Vector Machines with various kernel functions using an efficient set of arithmetic constraints.

Security Model. [48] consider two mutually distrusting parties – a malicious server and a semi-honest client, who follows the protocol but aims to learn information about the model's parameters. For their inference pipeline, they consider standard definitions of correctness, soundness, and zero-knowledge (similar to those by [56] and [33]). [48] do not provide a security definition for their proof of accuracy.

Discussion. Similar to Liu et al. [35], as [48] do not provide a security definition for their proof of accuracy, the precise security guarantee they achieve is somewhat unclear. However, [48] indicate that their scheme can be used to prove the accuracy on a public dataset, which falls within our framework of definition 3. This instantiation of their method is vulnerable to the same style of attack as outlined in §4.1.

Expanding/Restoring Security. Similar to Zhang et al. [56] and Liu et al. [35], the scheme of Wang and Hoang can be adapted using our techniques from § 5.3 to satisfy the guarantee with respect to freshly sampled datasets. In particular, their protocol can be plugged in as the ZKP of our protocol Π_{csp} in §5.2.

F Details of Secure CMC Framework

F.1 Auditing Protocol and Security Definitions

In this section, we formally define the security properties of an auditing protocol.

Completeness An auditing protocol Π is complete with error p if for any model h such that $F(h, S_{\text{train}}) = 1$, the following holds:

$$\Pr\left[b = 1 : (\mathsf{com}, b) \leftarrow \langle \mathsf{Prove}(h, S_{\text{train}}), \mathsf{Audit}\rangle\right] \geq 1 - p$$

Binding An auditing protocol Π is *computationally binding* if for any PPT adversary $\mathcal{A}$, the following is negligible in λ:

$$\Pr\left[\begin{array}{l} \mathsf{Commit}(h||S_{\text{train}};\rho) \\ = \mathsf{Commit}(h'||S'_{\text{train}};\rho') : (h, h', S_{\text{train}}, S'_{\text{train}}, \rho, \rho') \leftarrow \mathcal{A}(1^\lambda) \\ \wedge\ (h \neq h' \vee S_{\text{train}} \neq S'_{\text{train}}) \end{array}\right]$$

Note that we require the binding property to hold both for the model and the training dataset. This can be easily achieved by separately committing to the model and the training dataset with a standard binding commitment scheme (§ B.4), and then outputting the concatenation of the two commitments.

$\tilde{F}$-Relaxed Knowledge Soundness An auditing protocol Π is $\tilde{F}$-relaxed knowledge sound with knowledge error κ if for any PPT adversary Prove^*, there exists an expected polynomial time extractor $\mathsf{Ext}_{\mathsf{Prove}^*}$ such that the following holds:

$$p_{\text{ext}} \geq p_{\text{acc}} - \kappa$$

where

$$p_{\text{acc}} = \Pr\left[b = 1 : (\mathsf{com}, b) \leftarrow \langle \mathsf{Prove}^*, \mathsf{Audit}\rangle\right]$$

and p_{ext} is defined as

$$\Pr\left[\begin{array}{l} (\mathsf{com} = \mathsf{Commit}(h||S_{\text{train}};\rho) \wedge\ \tilde{F}(h, S_{\text{train}}) = 1) \\ \qquad\qquad \vee \\ (\mathsf{com} = \mathsf{Commit}(h||S_{\text{train}};\rho) = \mathsf{Commit}(h'||S'_{\text{train}}, \rho') \\ \qquad\qquad \wedge\ (h \neq h' \vee S_{\text{train}} \neq S'_{\text{train}})) \end{array}\right]$$

where $(\mathsf{com}, b) \leftarrow \langle \mathsf{Prove}^*, \mathsf{Audit}\rangle$ and $(h, S_{\text{train}}, \rho, h', S'_{\text{train}}, \rho') \leftarrow \mathsf{Ext}_{\mathsf{Prove}^*}(\mathsf{com})$,

Intuitively, this notion guarantees that if a cheating prover Prove^* convinces the auditor to accept with non-negligible probability, then it must either know a model h and a training dataset S_{train} satisfying a predicate $\tilde{F}$, **or** find two distinct openings to the same commitment com. As the latter event happens with negligible probability if Commit is computationally binding, this implies that Prove^* must know a valid model h and a training dataset S_{train} satisfying $\tilde{F}$. We call this property "$\tilde{F}$-relaxed" knowledge soundness because the predicate $\tilde{F}$ is a relaxation of the original predicate F. This is necessary because the auditor only checks the property f on a *finite sample*, which may not perfectly reflect the property F on the *underlying distribution*. In our concrete instantiations, we will quantify the gap between F and $\tilde{F}$ (see § F.2.1 and § F.2.2).

Zero Knowledge Let $\mathsf{view}_{\mathsf{Audit}}^{\mathsf{Prove}(h, S_{\text{train}})}$ be a string consisting of all the incoming messages that Audit receives from Prove during the interaction $\langle \mathsf{Prove}(h, S_{\text{train}}), \mathsf{Audit}\rangle$, and Audit's random coins. Π is *zero-knowledge* against semi-honest auditor if there exists a PPT simulator Sim such that for any PPT adversary $\mathcal{A}$, and any h such that $F(h, S_{\text{train}}) = 1$, the following is negligible in λ.

$$\left|\begin{array}{l} \Pr\left[b = 1 : b \leftarrow \mathcal{A}(\mathsf{view}_{\mathsf{Audit}}^{\mathsf{Prove}(h, S_{\text{train}})})\right] \\ -\Pr\left[b = 1 : \mathsf{view}' \leftarrow \mathsf{Sim}(1^\lambda);\quad b \leftarrow \mathcal{A}(\mathsf{view}')\right] \end{array}\right|$$

F.2 Construction of Auditing Protocol

We construct a commit-sample-prove auditing scheme Π_{csp}. While we focus on a protocol checking F on a hypothesis h only, the construction below can be naturally extended to a more complex F that additionally takes a training dataset as input. Let Commit be a binding commitment scheme (B.4) and $\mathsf{ZKP} = (\mathcal{P}, \mathcal{V})$ be a ZK proof system for the following relation $\mathcal{R}$: for a pair of public statement $\mathsf{x} = (\mathsf{com}, S_{\text{audit}})$ and private witness $\mathsf{w} = (h, \rho)$, we have $(\mathsf{x}, \mathsf{w}) \in \mathcal{R} \iff f(h, S_{\text{audit}}) = 1 \wedge \mathsf{com} = \mathsf{Commit}(h; \rho)$. We define a commit-sample-prove auditing protocol $\Pi_{\text{csp}} = (\mathsf{Commit}, \mathsf{Prove}, \mathsf{Audit})$ using an emprical predicate f and a distribution $\mathcal{D}$ over a query space $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^m$ as follows:

$\langle \mathsf{Prove}(h), \mathsf{Audit}\rangle$

1. Prove computes $\mathsf{com} = \mathsf{Commit}(h; \rho)$ using a uniformly random string $\rho \in \{0,1\}^{l_{\mathsf{Commit}}}$ and sends com to Audit.
2. Audit samples $S_{\text{audit}} \sim \mathcal{D}^n$ and sends it to Prove.
3. Prove and Audit execute $b \leftarrow \langle \mathcal{P}(\mathsf{w}), \mathcal{V}\rangle(\mathsf{x})$, where $\mathsf{x} = (\mathsf{com}, S_{\text{audit}})$ and $\mathsf{w} = (h, \rho)$. Here, Prove plays $\mathcal{P}$ and Audit plays $\mathcal{V}$.
4. Prove outputs com, while Audit outputs b.

We now state our main theorem regarding the security of Π_{csp}. The result is stated for a general auditing task defined by a predicate F and an empirical predicate f.

Theorem 3. *Suppose the empirical predicate f, the distributional predicate F, and the relaxed distributional predicate $\tilde{F}$ satisfy the following false negative and false positive rate bounds for every model h:*

$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) \neq 1 \mid F(h) = 1] \leq p_{fnr}$$

$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) = 1 \mid \tilde{F}(h) \neq 1] \leq p_{fpr}$$

Then Π_{csp} is a secure auditing protocol for F satisfying the following properties:

- *If* ZKP *is perfectly complete, then Π_{csp} is complete with error p_{fnr} for any model h such that $F(h) = 1$.*
- *If the underlying commitment scheme* Commit *is computationally binding, then Π_{csp} is computationally binding.*
- *If* ZKP *is knowledge sound with knowledge error κ, then Π_{csp} is $\tilde{F}$-relaxed knowledge sound with knowledge error $\kappa + p_{fpr}$.*
- *If p_{fnr} is negligible in λ[12], Commit is hiding, and ZKP is honest verifier zero-knowledge, then Π_{csp} is zero-knowledge against semi-honest auditor.*

Proof. Binding trivially follows from the computational binding property of the underlying commitment scheme.

Completeness: By the assumption on f and F, an honest auditing prover Prove fails to convince the auditor playing verifier $\mathcal{V}$ after receiving fresh S_{audit} with probability at most p_{fnr}. Since Commit and ZKP are perfectly complete, an honest ZKP prover $\mathcal{P}$ given any valid witness always convinces the verifier. Thus, the overall completeness error is at most p_{fnr}.

Knowledge Soundness: The protocol can be viewed as a commit-and-prove zero-knowledge proof with interleaved probabilistic tests on the statement. This approach is common in lattice-based zero-knowledge proofs (Cf. Theorem 5.1.6 of [37]).

Let Prove^* be any PPT adversary. We denote by $\mathcal{P}_1$ the interactive ZK prover algorithm that Prove^* invokes in Step 3 to prove the statement x fixed by the previous steps. Moreover, denote by $R_{\mathcal{E}}$ (resp. $R_{\mathcal{P}}$) the randomness space of the extractor $\mathcal{E}$ (resp. prover $\mathcal{P}_1$) for ZKP. We first construct the following extractor:

$\mathsf{Ext}_{\mathsf{Prove}^*}$

1. Run Prove^* to get com.
2. Sample $S \sim \mathcal{D}^n$, $r_{\mathcal{E}} \leftarrow R_{\mathcal{E}}$, and $r_{\mathcal{P}} \leftarrow R_{\mathcal{P}}$.
3. Let $\mathcal{P}_0$ be the algorithm that outputs $\mathsf{x} = (\mathsf{com}, S)$ as a statement and $\mathcal{P} = (\mathcal{P}_0, \mathcal{P}_1)$, where $\mathcal{P}_1$'s randomness is fixed to $r_{\mathcal{P}}$. Run $\mathcal{E}_{\mathcal{P}}(\mathsf{x}; r_{\mathcal{E}})$ to extract the witness (h, ρ).
4. If $f(h,S) \neq 1$ or $\mathsf{com} \neq \mathsf{Commit}(h;\rho)$, abort.
5. Repeat the following process:
   (a) Sample $S' \sim \mathcal{D}^n$ and $r'_{\mathcal{E}} \leftarrow R_{\mathcal{E}}$.
   (b) Let $\mathcal{P}'_0$ be the algorithm that outputs $\mathsf{x}' = (\mathsf{com}, S')$ as a statement and $\mathcal{P}' = (\mathcal{P}'_0, \mathcal{P}_1)$, where $\mathcal{P}_1$'s randomness is fixed to $r_{\mathcal{P}}$. Run $\mathcal{E}_{\mathcal{P}'}(\mathsf{x}'; r'_{\mathcal{E}})$ to extract the witness (h', ρ').
   (c) If $f(h', S') = 1$ and $\mathsf{com} = \mathsf{Commit}(h'; \rho')$, terminate and output (h, ρ, h', ρ')
   (d) Else, go to step (a).

Running time: Let T be the random variable counting the number of calls to the inner extractor $\mathcal{E}$ until termination. For each fixed com and prover's randomness $i \in R_{\mathcal{P}}$ we denote by ε_i the probability that $\mathcal{E}_{\mathcal{P}}((\mathsf{com}, S); r_{\mathcal{E}})$ successfully outputs (h, ρ) with $f(h,S) = 1$ and $\mathsf{com} = \mathsf{Commit}(h;\rho)$, where the probability is taken over $r_{\mathcal{E}} \leftarrow R_{\mathcal{E}}$ and $S \sim \mathcal{D}^n$. Denote by E the event that $\mathcal{E}_{\mathcal{P}}$ successfully outputs a valid witness at Step 3. We now evaluate the expected running time of Ext as follows:

$$
\begin{aligned}
\mathbb{E}[T] &= \sum_{i \in R_{\mathcal{P}}} \big(\mathbb{E}[T \mid r_{\mathcal{P}} = i \wedge E] \cdot \Pr[r_{\mathcal{P}} = i \wedge E] \\
&\qquad + \mathbb{E}[T \mid r_{\mathcal{P}} = i \wedge \neg E] \cdot \Pr[r_{\mathcal{P}} = i \wedge \neg E] \big) \\
&= \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \big(\mathbb{E}[T \mid r_{\mathcal{P}} = i \wedge E] \cdot \Pr[E \mid r_{\mathcal{P}} = i] \\
&\qquad + \mathbb{E}[T \mid r_{\mathcal{P}} = i \wedge \neg E] \cdot \Pr[\neg E \mid r_{\mathcal{P}} = i] \big) \\
&= \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \big(\mathbb{E}[T \mid r_{\mathcal{P}} = i \wedge E] \cdot \varepsilon_i \\
&\qquad + \mathbb{E}[T \mid r_{\mathcal{P}} = i \wedge \neg E] \cdot (1 - \varepsilon_i) \big) \\
&\leq \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \big(\tfrac{1}{\varepsilon_i} \cdot \varepsilon_i + 1 \cdot (1 - \varepsilon_i) \big) \\
&= 2
\end{aligned}
$$

where we used the fact that $\mathbb{E}[T \mid r_{\mathcal{P}} = i \wedge \neg E] = 1$ since the algorithm terminates after the initial extraction fails. Moreover, since the underlying ZKP is knowledge sound, each call to $\mathcal{E}$ runs in expected polynomial time. Overall, we conclude that Ext runs in expected polynomial time.

Knowledge error: We define the following events:

- E: $\mathcal{E}_{\mathcal{P}}(\mathsf{x}; r_{\mathcal{E}})$ succeeds at Step 3
- E': $\mathsf{Ext}_{\mathsf{Prove}^*}$ outputs valid (h, ρ, h', ρ')
- E_1: $E \wedge E'$
- E_2: $h = h'$
- E_3: $\tilde{F}(h) = 1$

Our goal is to relate the success probability p_{ext} of the extractor to p_{acc}, where

$$
\begin{aligned}
p_{\text{ext}} &\geq \Pr[E_1 \wedge (\neg E_2 \vee E_3)] \\
p_{\text{acc}} &= \Pr\left[b = 1 : (\mathsf{com}, b) \leftarrow \langle \mathsf{Prove}^*, \mathsf{Audit} \rangle\right]
\end{aligned}
$$

[12] We require this to realize negligible distinguishing advantage as in the standard zero-knowledge definition. If p_{fnr} is small but not negligible, one could relax the zero-knowledge property for CMC to allow for non-negligible distinguishing advantage. This relaxation is analogous to knowledge soundness with potentially non-negligible knowledge error including p_{fpr}.

To this end, we first rewrite (the lower bound of) p_{ext} as follows:

$$p_{\text{ext}} \geq \Pr[E_1 \wedge (\neg E_2 \vee E_3)] = \Pr[E_1] - \Pr[E_1 \wedge E_2 \wedge \neg E_3]$$

We now bound $\Pr[E_1]$ and $\Pr[E_1 \wedge E_2 \wedge \neg E_3]$ separately.
Bounding $\Pr[E_1]$**:** We rewrite $\Pr[E_1]$ as follows:

$$\Pr[E_1] = \Pr[E' \mid E] \cdot \Pr[E] \geq p_{\text{acc}} - \kappa$$

where $\Pr[E' \mid E] = 1$ since if E happens, then $\varepsilon_i \neq 0$ and the extractor always terminates in expected polynomial time and outputs (h, ρ, h', ρ') at Step 5. The last inequality follows from the definition of p_{acc} and the knowledge soundness of ZKP.

Bounding $\Pr[E_1 \wedge E_2 \wedge \neg E_3]$**:** For each fixed com and $i \in R_{\mathcal{P}}$, let $\mathcal{G}_i$ be the set of "good" pairs $(r_{\mathcal{E}}, S)$ that cause $\mathcal{E}_{\mathcal{P}}((\mathsf{com}, S); r_{\mathcal{E}})$ to output a valid witness (h, ρ) such that $f(h, S) = 1$ and $\mathsf{com} = \mathsf{Commit}(h; \rho)$. Then we have that

$$\begin{aligned}\varepsilon_i &= \Pr_{(r_{\mathcal{E}}, S) \leftarrow R_{\mathcal{E}} \times \mathcal{D}^n}[(r_{\mathcal{E}}, S) \in \mathcal{G}_i] = \Pr[E \mid r_{\mathcal{P}} = i] \\ &= \Pr_{(r'_{\mathcal{E}}, S') \leftarrow R_{\mathcal{E}} \times \mathcal{D}^n}\left[(r'_{\mathcal{E}}, S') \in \mathcal{G}_i \mid r_{\mathcal{P}} = i \wedge E\right].\end{aligned}$$

Given this, we rewrite $\Pr[E_1 \wedge E_2 \wedge \neg E_3]$ as follows:

$$\begin{aligned}
&\Pr[E_1 \wedge E_2 \wedge \neg E_3] \\
&= \sum_{i \in R_{\mathcal{P}}} \Pr[r_{\mathcal{P}} = i] \cdot \Pr[E \wedge E' \wedge E_2 \wedge \neg E_3 | r_{\mathcal{P}} = i] \\
&= \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \Pr[E | r_{\mathcal{P}} = i] \cdot \Pr[E' \wedge E_2 \wedge \neg E_3 | r_{\mathcal{P}} = i \wedge E] \\
&= \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \varepsilon_i \cdot \Pr[E' \wedge E_2 \wedge \neg E_3 | r_{\mathcal{P}} = i \wedge E] \\
&\leq \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \varepsilon_i \cdot \Pr\left[f(h, S') = 1 \wedge \tilde{F}(h) \neq 1 \,\middle|\, \begin{matrix} r_{\mathcal{P}} = i \wedge E \\ \wedge (r'_{\mathcal{E}}, S') \in \mathcal{G}_i \end{matrix}\right] \\
&\leq \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \varepsilon_i \cdot \frac{\Pr\left[f(h, S') = 1 \wedge \tilde{F}(h) \neq 1 \,\middle|\, r_{\mathcal{P}} = i \wedge E\right]}{\Pr\left[(r'_{\mathcal{E}}, S') \in \mathcal{G}_i \,\middle|\, r_{\mathcal{P}} = i \wedge E\right]} \\
&= \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \Pr\left[f(h, S') = 1 \wedge \tilde{F}(h) \neq 1 \,\middle|\, r_{\mathcal{P}} = i \wedge E\right] \\
&\leq \frac{1}{|R_{\mathcal{P}}|} \sum_{i \in R_{\mathcal{P}}} \Pr\left[f(h, S') = 1 \,\middle|\, r_{\mathcal{P}} = i \wedge E \wedge \tilde{F}(h) \neq 1\right] \\
&\leq p_{\text{fpr}}
\end{aligned}$$

where the last inequality follows from the assumption on f and $\tilde{F}$, as at this stage h is fixed by event E and elements in each element of S' is sampled i.i.d from $\mathcal{D}^n$.

Combining the bounds, we get

$$\begin{aligned}p_{\text{ext}} &\geq \Pr[E_1] - \Pr[E_1 \wedge E_2 \wedge \neg E_3] \\ &\geq p_{\text{acc}} - \kappa - p_{\text{fpr}}\end{aligned}$$

This completes the proof.

Zero Knowledge: We construct the following simulator, internally using the simulator $\mathcal{S}$ for ZKP:
$\underline{\mathsf{Sim}(1^\lambda)}$

1. Generate a dummy commitment $\mathsf{com} \leftarrow \mathsf{Commit}(0; \rho)$ using a uniformly random string $\rho \in \{0,1\}^{\ell_\rho}$.
2. Sample $S_{\text{audit}} \sim \mathcal{D}^n$.
3. Run $\mathcal{S}((\mathsf{com}, S_{\text{audit}}))$ to get a simulated view' for ZKP.
4. Output $\mathsf{view} = (\mathsf{com}, S_{\text{audit}}, \mathsf{view}')$.

Since Commit is hiding, the dummy commitment is indistinguishable from a real commitment. Moreover, the output of $\mathcal{S}$ is indistinguishable from the view of Audit during the interaction $\langle \mathcal{P}(h), \mathcal{V} \rangle((\mathsf{com}, S_{\text{audit}}))$ for any valid witness h. Since a real execution of Π_{csp} defines $\mathsf{x} = (\mathsf{com}, S_{\text{audit}})$ and $\mathsf{w} = h$ such that $\mathcal{R}(\mathsf{x}, \mathsf{w}) = 1$ except with probability at most p_{fnr}, by setting p_{fnr} to be negligible in λ, the output of Sim is indistinguishable from the view of Audit during the interaction $\langle \mathsf{Prove}(h), \mathsf{Audit} \rangle$. This completes the proof. □

F.2.1 Example Instantiation: Π_{csp} for Accuracy Auditing

To instantiate Π_{csp} for accuracy auditing, we consider the empirical and true error rates as follows:

$$\begin{aligned}\hat{\ell}_S(h) &= \frac{1}{n} \sum_{(x,y) \in S} \mathbb{I}(h(x) \neq y) \\ \ell(h) &= \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{I}(h(x) \neq y)]\end{aligned}$$

where $n = |S|$, and define the empirical predicate f, the distributional predicate F, and the relaxed predicate $\tilde{F}$ as follows.

$$\begin{aligned}f(h, S_{\text{audit}}) = 1 &\iff \hat{\ell}_S(h) \leq t + \delta \\ F(h) = 1 &\iff \ell(h) \leq t \\ \tilde{F}(h) = 1 &\iff \ell(h) \leq t + 2\delta\end{aligned}$$

To apply Theorem 3 to accuracy auditing, it would be sufficient to find the false negative rate p_{fnr} and false positive rate p_{fpr} by the following lemma, and then set $n\delta^2 \in \Omega(\lambda)$.

Lemma 3. *For any hypothesis h,*

$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) \neq 1 \mid F(h) = 1] \leq 2e^{-2n\delta^2} \quad (1)$$

$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) = 1 \mid \tilde{F}(h) \neq 1] \leq 2e^{-2n\delta^2} \quad (2)$$

Proof. Consider the empirical error $\hat{\ell}_S(h) = \frac{1}{n}\sum_{(x,y) \in S} \mathbb{I}(h(x) \neq y)$ and the true error $\ell(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{I}(h(x) \neq y)]$. Since each element of S_{audit} is sampled i.i.d. from $\mathcal{D}$, by Hoeffding's inequality, we have

$$\Pr_{S_{\text{audit}} \leftarrow \mathcal{D}^n}[|\hat{\ell}_S(h) - \ell(h)| \geq \delta] \leq 2e^{-2n\delta^2}.$$

Thus, for any h such that $F(h) = 1$ (i.e., $\ell(h) \le t$), the probability that $f(h, S_{\text{audit}}) \neq 1$ (i.e., $\hat{\ell}_S(h) > t + \delta$) is at most $2e^{-2n\delta^2}$.

Similarly, for any h such that $\tilde{F}(h) \neq 1$ (i.e., $\ell(h) > t + 2\delta$), the probability that $f(h, S_{\text{audit}}) = 1$ (i.e., $\hat{\ell}_S(h) \le t + \delta$) is at most $2e^{-2n\delta^2}$. □

Overall, by Theorem 3 and Lemma 3, we conclude that Π_{csp} instantiated for accuracy auditing is a secure auditing protocol for accuracy with the following properties for a sufficiently large $n\delta^2 \in \Omega(\lambda)$:

- If the true error of the model h is $\le t$, then the auditor accepts with high probability.
- If the auditor accepts, then the auditor gets assurance that the true error of the model h is at most $t + 2\delta$, even if the prover is misbehaving.

F.2.2 Example Instantiation: Π_{csp} for Demographic Parity Auditing

Similarly, we can instantiate the protocol Π_{csp} for auditing fairness conditions. Take the demographic parity as an example, for which we can set the query space to $Q = \{x_i\}_{i=1}^m$ without ground-truth labels. We consider the empirical and true demographic parity differences as follows:

$$\Delta_{\mathsf{dp}}(h, S_{\text{audit}}) = \left| \frac{1}{n_0} \sum_{x \in S_0} \mathbb{I}(h(x) = 1) - \frac{1}{n_1} \sum_{x \in S_1} \mathbb{I}(h(x) = 1) \right|$$

$$\hat{\Delta}_{\mathsf{dp}}(h) = |\mathrm{E}_{x \sim \mathcal{D}}[\mathbb{I}(h(x) = 1) \,|\, s_x = 0] - \mathrm{E}_{x \sim \mathcal{D}}[\mathbb{I}(h(x) = 1) \,|\, s_x = 1]|$$

where s_x denotes the sensitive feature of a data point x, $S_0 = \{x \in S_{\text{audit}} : s_x = 0\}$, $S_1 = \{x \in S_{\text{audit}} : s_x = 1\}$, $n_0 = |S_0|$, and $n_1 = |S_1|$.

Define the empirical predicate f, the distributional predicate F, and the relaxed predicate $\tilde{F}$ as follows.

$$f(h, S_{\text{audit}}) = 1 \iff \Delta_{\mathsf{dp}}(h, S_{\text{audit}}) \le t + 2\delta$$
$$F(h) = 1 \iff \hat{\Delta}_{\mathsf{dp}}(h) \le t$$
$$\tilde{F}(h) = 1 \iff \hat{\Delta}_{\mathsf{dp}}(h) \le t + 4\delta$$

To apply Theorem 3 to demographic parity auditing, we can prove the following lemma in place of Lemma 3, and then set $n_{\min}\delta^2 \in \Omega(\lambda)$.

Lemma 4. *For any hypothesis h,*

$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) \neq 1 \mid F(h) = 1] \le 4e^{-2n_{\min}\delta^2} \quad (3)$$

$$\Pr_{S_{audit} \sim \mathcal{D}^n}[f(h, S_{audit}) = 1 \mid \tilde{F}(h) \neq 1] \le 4e^{-2n_{\min}\delta^2} \quad (4)$$

where $n_{\min} = \min(n_0, n_1)$.

Proof. We first prove (4). Define the following variables:

$$g_0 = \frac{1}{n_0} \sum_{x \in S_0} \mathbb{I}(h(x) = 1) \quad p_0 = \mathrm{E}_{x \sim \mathcal{D}}[\mathbb{I}(h(x) = 1) \,|\, s_x = 0]$$

$$g_1 = \frac{1}{n_1} \sum_{x \in S_1} \mathbb{I}(h(x) = 1) \quad p_1 = \mathrm{E}_{x \sim \mathcal{D}}[\mathbb{I}(h(x) = 1) \,|\, s_x = 1]$$

By Hoeffding's inequality, we have

$$\Pr[|g_0 - p_0| \ge \delta] \le 2e^{-2n_0\delta^2}$$
$$\Pr[|g_1 - p_1| \ge \delta] \le 2e^{-2n_1\delta^2}$$

where the probability is taken over the randomness of $S_{\text{audit}} \sim \mathcal{D}^n$. Thus, for any h such that $\tilde{F}(h) \neq 1$ (i.e., $|p_0 - p_1| > t + 4\delta$), the probability that $f(h, S_{\text{audit}}) = 1$ (i.e., $|g_0 - g_1| \le t + 2\delta$) can be bounded as follows:

$$\begin{aligned}
\Pr[|g_0 - g_1| \le t + 2\delta] &\le \Pr[|g_0 - g_1| \le |p_0 - p_1| - 2\delta] \\
&\le \Pr[2\delta \le |p_0 - p_1 - (g_0 - g_1)|] \\
&\le \Pr[2\delta \le |p_0 - g_0| + |p_1 - g_1|] \\
&\le \Pr[\delta \le |p_0 - g_0| \vee \delta \le |p_1 - g_1|] \\
&\le \Pr[\delta \le |p_0 - g_0|] + \Pr[\delta \le |p_1 - g_1|] \\
&\le 2e^{-2n_0\delta^2} + 2e^{-2n_1\delta^2} \le 4e^{-2n_{\min}\delta^2}
\end{aligned}$$

Analogously, we can prove (3). For any h such that $F(h) = 1$ (i.e., $|p_0 - p_1| \le t$), the probability that $f(h, S_{\text{audit}}) \neq 1$ (i.e., $|g_0 - g_1| > t + 2\delta$) can be bounded in the same way:

$$\begin{aligned}
&\Pr[|g_0 - g_1| > t + 2\delta] \\
&\le \Pr[|g_0 - g_1| > |p_0 - p_1| + 2\delta] \le 4e^{-2n_{\min}\delta^2}
\end{aligned}$$

□

Overall, by Theorem 3 and Lemma 4, we conclude that Π_{csp} instantiated for demographic parity auditing is a secure auditing protocol for demographic parity with the following properties for a sufficiently large $n_{\min}\delta^2 \in \Omega(\lambda)$:

- If the true demographic parity of the model h is $\le t$, then the auditor accepts with high probability.
- If the auditor accepts, then the auditor gets assurance that the true demographic parity of the model h is at most $t + 4\delta$, even if the prover is misbehaving.

G Further Evaluation

First, we present in Figure 5 results for attacking accuracy audits on additional datasets mentioned in §7.

Next, we present the application of the attack described in § 4.1. In this attack, the adversary is attempting to maximize the model's denial rate $\Pr_{x \sim \mathcal{D}}[h(x) = 0]$ while still appearing

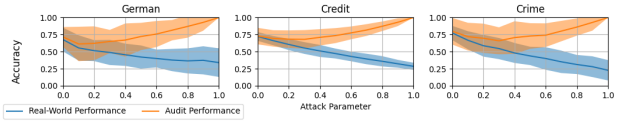

Figure 5: Accuracy of models trained on datasets constructed by Algorithm 1 on various benchmarks. Values are averages over ten runs, error bars represent one standard deviation.

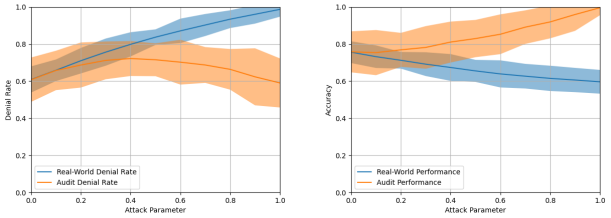

Figure 6: Accuracy and denial rates of models trained on datasets constructed by Algorithm 1 on ACSEmployment. Values are averages over ten runs, error bars represent one standard deviation.

accurate to the audit. The results of this attack are given in Figure 6. Observe that as the attack parameter approaches 1 (and the attack becomes maximally malicious), the denial rate of the model on the audit set remains close to the fully honest denial rate while the denial rate on independently sampled data approaches 1. Similarly, the accuracy of the model on the audit set approaches 1, while the true accuracy decreases down to roughly 0.6 (this reflects the true denial rate of the distribution).

G.1 Attacking XGBoost

We next show that a minor modification of Algorithm 1 suffices to evade accuracy auditing of XGBoost models [14]. Unlike a single decision tree, XGBoost trains a gradient-boosted ensemble of trees, so the theoretical guarantee of Theorem 1 does not apply: with only one copy of S_{audit} in S'_{train}, the boosting process assigns excessive weight to the $2d$ flipped-label neighborhood points, causing the ensemble to misclassify the audit points themselves and thereby fail the audit. To counteract this, we include $2d$ copies of S_{audit} (with true labels) in S'_{train} rather than a single copy, so that the correct-label signal at each audit point outweighs its $2d$ flipped-label neighbors; all other details of the construction remain unchanged. We used the perturbation magnitude $\varepsilon = 1.0$ throughout. The XGBoost ensemble is configured with 500 trees of maximum depth 10, learning rate $\eta = 0.3$, L2 regularization $\lambda = 1.0$, and minimum split-loss $\gamma = 0$; all other XGBoost hyperparameters are left at their defaults. We empirically evaluated this attack on the Default Credit, Adult Income, and ACS Income datasets; results are shown in Figure 7. Our modified attack

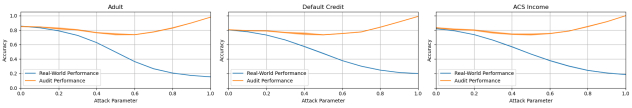

Figure 7: Accuracy of XGBoost models trained on inflated datasets constructed by the modified Algorithm 1 on Adult, Default Credit, and ACS Income datasets. Values are averages over five runs; error bars represent one standard deviation.

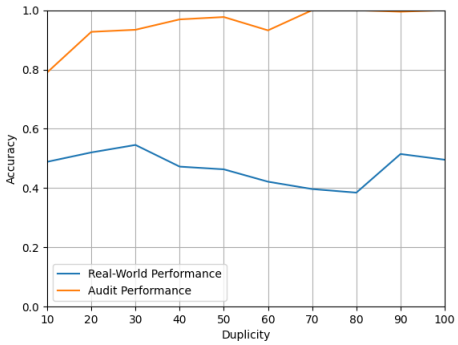

Figure 8: Performance of 226M-parameter neural networks trained on datasets constructed from a mixture of Gaussian distributions. Duplicity refers to the number of perturbed copies of the audit dataset included in the training data.

successfully evades accuracy auditing across all benchmarks.

G.2 Attacking Neural Networks

Finally, we present an evaluation of a modified version of the attack that targets neural networks rather than decision trees. Whereas decision trees have very specific conditions that allow us to constrain their behavior, it is much harder to provide theoretical guarantees for neural networks. In order to encourage memorization of the training data, we used a relatively shallow network with very large individual layers. Our attack samples a large amount of training data, and decides whether to label each point with the honest label or dishonest label depending on its proximity to the nearest audit data point. We evaluated this attack on an 8-dimensional mixture of Gaussian distributions; the results are shown in Figure 8.

We also performed this attack on ACS Employment data, and over 10 runs, trained a model with average accuracy of 69.83% accuracy on held-out test data while maintaining an average of 88.82% accuracy on audit data.

G.3 Evading Statistical Methods for Attack Detection

We empirically examine our ability to observe data forging attacks by applying statistical tests to the datasets, as described in § D. There is no singular way to determine whether two samples were drawn from the same distribution, so we apply some common statistical tools. In particular, our goal is to determine if the distribution from which the audit data is drawn is identical to the distribution from which the training data is drawn. We use Welch's t-test, which serves to determine whether two distributions have the same mean, and Levene's test, a one-way ANOVA for determining whether two distributions have the same variance. These tests are typically applied to 1-dimensional data, and so we apply them to each feature individually. The results of these experiments are given in Table 2.

We observe that the summary statistics of the malicious training data closely match the values for the honest data, suggesting that comparing these two values would not be a successful detection mechanism. This is compounded by the fact that the test statistics for Welch's t-test and Levene's test for the malicious training data are considerably smaller on average than the same test statistics for the honest training data, corroborating higher rate of passing the hypothesis tests we observe. At a significance level of $\alpha = 0.05$, we expect a false positive rate of approximately 5%. On the other hand, we observe a 0% true positive rate. We note that in a practical application of this attack, the auditor would have access only to the honest or malicious values over a single training run, and would thus be unable to easily distinguish between the two cases by comparing the values or by looking at averages over many runs as we have done here. That being said, an auditor may find it suspicious if the p-value returned by a statistical test is extremely low (even though such a scenario may be very plausible for some distributions); an attacker can safely relax this attack to a comfortable degree, though doing so will increase the risk of failing the audit.

Table 2: Summary and Test statistics for Age feature on ACSEmployment, conditioned on label. Test statistics used are Welch's t-test and Levene's test. Attack is undetectable when summary statistics are similar to honest ones, and when test statistics are close to 0. Comparisons are between fully honest and fully malicious datasets.

Age		Label = 0		Label = 1	
		Honest	Attack	Honest	Attack
Summary Statistics	μ	41.6651	41.9657	43.9184	43.8131
	σ^2	804.5804	810.8822	223.1269	221.42394
Test Statistics	t-test	0.6521	0.0033	0.7067	0.0110
	ANOVA	0.6200	0.0026	1.6500	0.0186

Education		Label = 0		Label = 1	
		Honest	Attack	Honest	Attack
Summary Statistics	μ	13.39761692	13.41700338	18.45539675	18.50545506
	σ^2	42.99789908	42.16899485	9.979327135	8.943082831
Test Statistics	t-test	0.7984001575	0.0356390553	0.9499974697	0.1302788154
	ANOVA	0.4844657261	0.0003374130653	1.227829625	0.02531893152

Military Status		Label = 0		Label = 1	
		Honest	Attack	Honest	Attack
Summary Statistics	μ	2.5794	2.5834	3.8121	3.8302
	σ^2	3.2749	3.2648	0.3507	0.3265
Test Statistics	t-test	0.4997	0.0313	0.8699	0.1755
	ANOVA	1.0240	0.0009	1.2394	0.0304