ROBUST FEDERATED INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated inference, in the form of one-shot federated learning, edge ensembles, or federated ensembles, has emerged as an attractive solution to combine predictions from multiple models. This paradigm enables each model to remain local and proprietary while a central server queries them and aggregates predictions. Yet, the robustness of federated inference has been largely neglected, leaving them vulnerable to even simple attacks. To address this critical gap, we formalize the problem of robust federated inference and provide the first robustness analysis of this class of methods. Our analysis of averaging-based aggregators shows that the error of the aggregator is small either when the dissimilarity between honest responses is small or the margin between the two most probable classes is large. Moving beyond linear averaging, we show that problem of robust federated inference with non-linear aggregators can be cast as an adversarial machine learning problem. We then introduce an advanced technique using the DeepSet aggregation model, proposing a novel composition of adversarial training and test-time robust aggregation to robustify non-linear aggregators. Our composition yields significant improvements, surpassing existing robust aggregation methods by $4.7 - 22.2\%$ in accuracy points across diverse benchmarks.

1 INTRODUCTION

Over the past several years, concepts such as *one-shot federated learning* (OFL) (Dai et al., 2024; Diao et al., 2023; Zhang et al., 2022; Guha et al., 2019), *edge ensembles* (Malka et al., 2025; Shlezinger et al., 2021), and *federated ensembles* (Allouah et al., 2024a; Hamer et al., 2020) have gained traction in collaboratively performing inference from several client-local models. More recently, the availability of diverse open-source large language models (LLMs) has spurred interest in aggregating outputs from multiple models to answer a given query, giving rise to sophisticated *LLM ensembles* that leverage complementary strengths of individual models (Tekin et al., 2024; Wang et al., 2024; Jiang et al., 2023). Despite being introduced under different names, these techniques share a common principle: combining predictions from multiple client-held models to produce a single output. In this paper, we collectively refer to these approaches under the terminology of *federated inference*. In this setting, clients retain proprietary (locally trained) models, while a central server queries them for inference as illustrated in Figure 1. The individual predictions are then aggregated into a final prediction, either using averaging-based aggregations (Dai et al., 2024; Zhang et al., 2022; Gong et al., 2022) or server-side aggregator neural networks (Allouah et al., 2024a; Wang et al., 2024).

While federated inference is gaining traction, its robustness to model failures and poisoned outputs remains largely overlooked in the literature, despite some initial preliminary study (Liu et al., 2022). This gap is critical for two reasons: *(i)* failures and errors are practically unavoidable, and *(ii)* existing work in robust statistics and Byzantine-robust machine learning (ML) demonstrates that undefended models are inherently vulnerable, even to relatively simple attacks (Guerraoui et al., 2024; Diakonikolas & Kane, 2023). It is therefore of paramount importance to clearly define the potential threats that may arise in federated inference, so as to prevent a technological advantage from becoming a significant vulnerability. In this paper, we take a step toward closing this gap by defining potential failure modes of federated inference and presenting the first robustness analysis of this class of schemes.

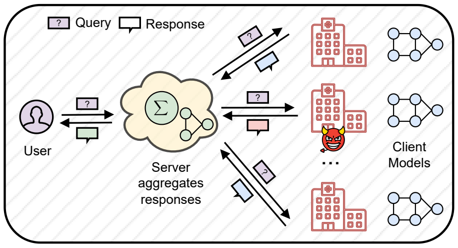

Figure 1: Federated Inference

DeepSet	CWTM	Adv. Tr.	CIFAR-10	CIFAR-100	AG-News
✓	✗	✗	46.0 ± 3.9	47.4 ± 3.1	76.4 ± 1.9
✓	✓	✗	47.0 ± 3.8	67.0 ± 0.7	76.7 ± 2.2
✓	✗	✓	48.6 ± 6.7	65.1 ± 0.8	76.7 ± 0.9
✓	✓	✓	$\mathbf{51.4 \pm 2.2}$	$\mathbf{68.0 \pm 0.8}$	$\mathbf{77.5 \pm 1.2}$

Table 1: Evaluation of robust elements in a setup with $n = 17, f = 4$. We report the worst-case accuracy across 5 different attacks (see Appendix E.4 for details).

1.1 Main Contributions

Problem formulation and analysis of the averaging aggregator. We formalize the problem of robust federated inference for the first time. We consider a system of n clients, each with a local data distribution and probit-valued classifier, where the outputs of up to $f < \frac{n}{2}$ clients may be arbitrarily corrupted at inference time. Our goal is to design aggregation schemes that remain accurate with respect to the global data distribution despite the corruptions. When the server uses an averaging-based aggregation (Dai et al., 2024; Zhang et al., 2022; Gong et al., 2022), a natural way to robustify is to substitute it with a robust averaging scheme (Allouah et al., 2023) which ensures that the output of the aggregator is an estimate of the average of the honest probits. Prominent examples include coordinate-wise trimmed mean (CWTM), coordinate-wise median (CWMed), *etc*. (Guerraoui et al., 2024). However, we show that robust averaging may prove insufficient since the aggregator's output can be sufficiently close to the honest average, yet produce a misclassification. Nevertheless, in the cases where averaging suffices, we derive formal robustness certificates for federated inference. Our analysis shows that the error of the aggregator depends upon the fraction of corruptions f/n, the margin between the top two classes, and the dissimilarity between the outputs of different clients.

Robust Inference as an adversarial ML problem. Beyond averaging, recent work has shown that non-linear trained aggregators (*e.g.*, neural networks) often outperform averaging-based aggregators in the uncorrupted setting (Allouah et al., 2024a; Wang et al., 2024). When using such aggregators, we show that the problem of robust inference can be cast as an adversarial learning problem over the probit-vectors. Encouragingly, contrary to the difficulty of standard adversarial ML problems in the image space (Goodfellow et al., 2014), we show that the adversarial problem on probit-vectors can be more reliably addressed, thanks to the structure of the input space where each corrupted vector is confined to a probability simplex over the number of classes. Yet, naively leveraging adversarial training (Madry et al., 2018) to solve the problem remains computationally intractable due to the high cardinality of permuting through different choices of adversarial clients at training time, since *any* f (unknown) clients out of n can be malicious.

Robust DeepSet Aggregator. To alleviate this issue, we propose to use a neural network aggregator based on the DeepSet model (Zaheer et al., 2017), an architecture which is invariant to the order of inputs. Specifically, since the aggregator's output can be independent to the order of clients, leveraging DeepSets enables us to reduce the search of choosing f adversaries to $\binom{n}{f}$ instead of requiring to permute through them $\binom{n}{f} f!$. In practice, we show that adversarial training by sampling any N choices of adversaries where $N \ll \binom{n}{f}$ suffices to achieve good performance with DeepSet. To further reduce the sensitivity of DeepSet to corrupted probits, we propose a composition of robust averaging with DeepSet at inference-time, which significantly boosts empirical performance. In particular, we show that this composition may only be applied at inference-time, preventing any escalation of training costs during adversarial training. By combining robust elements from both the adversarial ML and robust ML literature in our novel composition, we achieve the state-of-the-art (SOTA) performance for federated inference as illustrated in Table 1.

Empirical validation. To rigorously evaluate defenses, we design a new attack called the Strongest Inverted Attack (SIA) that challenges existing defenses. We conduct extensive experiments on three datasets (CIFAR-10, CIFAR-100, and AG-News) covering both vision and language modalities as well as diverse model families (ResNet-8, ViT-B/32, DistilBERT). Our approach yields a 4.7–22.2% points improvement over existing methods across a suite of 6 different attacks, including SIA.

1.2 Related Work

Federated Inference. Collaborative inference through client ensembles has been explored in one-shot Federated Learning (FL) (Guha et al., 2019; Gong et al., 2022; Dai et al., 2024) and decentralized edge networks (Shlezinger et al., 2021; Malka et al., 2025), motivated by communication efficiency, reduced training cost, or proprietary nature of client models accessible only via black-box inference. Another common scenario is when clients already possess trained models which are offered in model market scenarios (Li et al., 2021). Yet, the robustness of federated inference has received limited attention. An initial preliminary work was conducted by Liu et al. (2022) in the vertical federated learning setting, where they try to reconstruct the underlying uncorrupted responses using an autoencoder and a block-sparse optimization process. Their method, CoPur, purifies responses before aggregation on the server. However, purification alone is vulnerable to stronger attacks and highly sensitive to input magnitudes, since it operates in the logit space rather than probits.

Byzantine Distributed Learning. Byzantine robustness has become a central topic in distributed ML (Alistarh et al., 2018; Chen et al., 2017; Farhadkhani et al., 2022; Guerraoui et al., 2024), where malicious clients may send arbitrary updates during training. Our setting differs in focusing on inference rather than training, and in not assuming fixed Byzantine identities (*cf.*, (Dorfman et al., 2024)). While the objectives thus diverge, a key idea in Byzantine distributed learning, namely robust averaging, remains relevant to our analysis. Indeed, robust averaging techniques can be adapted to improve the resilience of federated inference in the presence of corrupted client outputs, at the cost of a some technicalities, as demonstrated in Section 3.

Federated Distillation and Robust Voting. In FL, many works propose to share logits on a public dataset instead of gradients to improve communication efficiency and privacy (Fan et al., 2023; Gong et al., 2022; Sattler et al., 2021). Recent work addresses adversarial logits by proposing robust aggregation methods (Roux et al., 2025; Li et al., 2024; Mi et al., 2021), but these approaches rely on assumptions inapplicable to our setting. For example, ExpGuard (Roux et al., 2025) tracks client behavior across rounds, whereas in our case clients may act arbitrarily at each inference. Similarly, FedMDR (Mi et al., 2021) requires clients to report accuracy on a public dataset to weight contributions, information unavailable in our setup. A parallel line of work studies robust voting, where voters/clients cast scores that are aggregated to resist malicious participants, either in general settings (Allouah et al., 2024b) or in federated learning specifically (Chu & Laoutaris, 2024; Cao et al., 2022). These approaches, however, typically assume fixed client identities or some control over the training process, assumptions which do not hold in our setting.

2 Problem of Robust Federated Inference

We consider a classification task, mapping an input space $\mathcal{X}$ to an output space $\mathcal{Y} = [K] := \{1, \ldots, K\}$, and a system comprising n clients, each with a local data generating distribution $\mathcal{D}_i$ over $\mathcal{X} \times \mathcal{Y}$. For each client $i \in [n]$, we are given a probit-valued regressor h_i mapping $\mathcal{X}$ to the simplex $\Delta^K := \{z \in [0,1]^K \mid \sum_{k \in [K]} z_k = 1\}$, which plays the role of a local classifier. In this context, the goal of a federated inference scheme is to design a mapping $\psi : \left(\Delta^K\right)^n \to [K]$ that aggregates clients' local probits in order to minimize the expected prediction error on the mixture of distributions $\mathcal{D} := \frac{1}{n}\sum_{i=1}^n \mathcal{D}_i$. Specifically, denoting $\mathbf{h}(x) := (h_1(x), \ldots, h_n(x))$, we seek an aggregator ψ_{o}, from a set of candidate aggregators Ψ, that minimizes the *federated inference risk*

$$\mathcal{R}(\psi) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_\psi(x,y)\right], \quad \text{where} \quad \ell_\psi(x,y) := \mathbb{1}\left\{\psi\left(\mathbf{h}(x)\right) \neq y\right\}. \tag{1}$$

A typical example of aggregation is $\psi(\mathbf{h}(x)) := \arg\max_{k\in[K]} \left[\frac{1}{n}\sum_{i=1}^n h_i(x)\right]_k$, where $[\cdot]_k$ denotes the k-th coordinate of the vector and where the $\arg\max$ breaks ties arbitrarily. More generally, however, Ψ may involve non-linear aggregation schemes. In fact, prior work (Allouah et al., 2024a; Wang et al., 2024) has shown that such aggregation strategies yield higher accuracy expectation, compared to any individual classifier $h_1, \ldots, h_n$.

Robust federated inference. We are interested in solving the problem of *robust federated inference*, wherein a fraction of the clients' is subject to corruption prior to aggregation. Specifically, we consider a scenario wherein for each query $x \in \mathcal{X}$, up to $f < n/2$ of the n clients (of hidden identity) can return arbitrarily corrupted vectors in the probit space Δ^K. Our goal is to design an aggregation scheme that yields high accuracy, despite such corruptions. In what follows, for any input $x \in \mathcal{X}$, we denote by $\Gamma_f(x)$ the set of all possible probits after up to f corruptions, i.e.,

$$\Gamma_f(x) \coloneqq \left\{ \mathbf{z} = (z_1, \ldots, z_n) \in (\Delta^K)^n \,\middle|\, \exists\, H \subseteq [n],\ |H| \geq n - f,\ \forall i \in H,\ z_i = h_i(x) \right\}. \quad (2)$$

Formally, we seek an aggregator $\psi_{\text{rob}} \in \Psi$ minimizing the *robust federated inference risk*, given by

$$\mathcal{R}_{\text{adv}}(\psi) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\psi}^{\text{adv}}(x, y) \right], \quad \text{where} \quad \ell_{\psi}^{\text{adv}}(x, y) \coloneqq \max_{\mathbf{z} \in \Gamma_f(x)} \mathbb{1}\left\{ \psi(\mathbf{z}) \neq y \right\}. \quad (3)$$

In the following, we show that robust federated inference risk can be upper bounded by federated inference risk and an overhead resulting from probit corruptions. Thereby we show that if an optimal aggregator (with respect to the original federated inference risk (1)) is apriori known, then robust ensembling can be achieved by designing an aggregator that aims to minimize disagreement with the optimal aggregator, in the presence of corruptions. In doing so, consider an aggregation scheme ψ_{o} minimizing (1), i.e., the expected learning error without corruption. This aggregator ψ_{o} represents an *oracular aggregator*, i.e., optimal in the hypothetical scenario when we have access to the uncorrupted probits. Using ψ_{o} as reference for robustness, we can bound robust federated inference risk for an aggregator ψ_{rob} as per the following lemma (which we prove in Appendix A).

Lemma 1. *For any $x \in \mathcal{X}$, let $\hat{y}_o = \psi_o(\mathbf{h}(x))$. Then,*

$$\mathcal{R}_{adv}(\psi_{rob}) \leq \mathcal{R}(\psi_o) + \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\psi_{rob}}^{adv}(x, \hat{y}_o) \right]. \quad (4)$$

The overhead $\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\psi_{\text{rob}}}^{\text{adv}}(x, \hat{y}_{\text{o}}) \right]$, named the *robustness gap*, represents the excess error due to adversarial probit corruptions. This gap is the worst-case probability that ψ_{rob} disagrees with the oracular aggregator under up to f probits being corrupted. In the following, we analyze the robustness gap in the case when the oracular aggregator is given by the averaging operation and the robust aggregator satisfies the property of (f, κ)-robust averaging Allouah et al. (2023).[1]

3 Averaging as oracular aggregator

We first consider the case where the oracular aggregator ψ_{o} is based on computing the average of the uncorrupted probits, i.e.,

$$\psi_{\text{o}}(\mathbf{h}(x)) \coloneqq \operatorname*{argmax}_{k \in [K]} \left[\overline{h}(x) \right]_k, \quad \text{where } \overline{h}(x) \coloneqq \frac{1}{n} \sum_{i=1}^{n} h_i(x)$$

In this particular case, we robustify the aggregation against probit corruptions by substituting the averaging by a robust averaging $\textsc{RobAvg} : (\Delta^K)^n \to \mathbb{R}^K$, taking inspiration from the literature of Byzantine-robust machine learning (Guerraoui et al., 2024) and robust mean estimation (Diakonikolas & Kane, 2023). Specifically, we set

$$\psi_{\text{rob}}(\mathbf{z}) \coloneqq \operatorname*{argmax}_{k \in [K]} \left[\textsc{RobAvg}\,(\mathbf{z}) \right]_k. \quad (5)$$

Where RobAvg is a robust averaging aggregation rule. The concept of robust averaging, initially introduced by Allouah et al. (2023) can be defined as follows.

Definition 1 (Robust averaging). *Let $\kappa \geq 0$. An aggregation rule* RobAvg *is (f, κ)*-robust *if, for any set of n vectors $v_1, ..., v_n \in \mathbb{R}^d$ and any set $S \subseteq [n]$ of size $n - f$, the following holds true:*

$$\left\| \textsc{RobAvg}(v_1, ..., v_n) - \overline{v}_S \right\|_2^2 \leq \frac{\kappa}{|S|} \sum_{i \in S} \left\| v_i - \overline{v}_S \right\|_2^2,$$

where $\overline{v}_S \coloneqq \frac{1}{n-f} \sum_{i \in S} v_i$, and parameter κ is called the robustness coefficient

Essentially, robust averaging ensures that despite up to f inputs being arbitrarily corrupted the output of the aggregator is an estimate of the uncorrupted vectors' average. The estimation error is bounded by the "empirical variance" of the uncorrupted input vectors, times a constant value κ.

[1] The analysis can be easily extended to weighted averaging by simply re-scaling the inputs to the aggregator.

Examples of robust averaging include CWTM, CWMed and geometric median (GM) (Guerraoui et al., 2024, Chapter 4). Specifically, CWTM is (f, κ)-robust with $\kappa = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$. While robust averaging ensures proximity to the average of uncorrupted input vectors in ℓ_2-norm, even a small estimation error can result in large prediction error due to the non-continuity of the argmax_k operator. We illustrate this insufficiency of robust averaging through the following counter example.

Counter example. Let $\varepsilon > 0$. Consider a point $x \in \mathcal{X}$ and three possible probit vectors for x, $h_1(x), h_2(x), h_3(x) \in \Delta^3$ defined as follows:

$$h_1(x) = (1, 0, 0), \quad h_2(x) = (0, 1, 0), \quad \text{and } h_3(x) = (1/2, 1/2 - \varepsilon, \varepsilon).$$

Note that $\overline{h}(x) = (1/2, 1/2 - \varepsilon, \varepsilon)$. Now, consider $\hat{v} = (1/2 - \varepsilon, 1/2, \varepsilon)$. We have $\left\|\overline{h}(x) - \hat{v}\right\|_2 = \left(2\varepsilon^2\right)^{1/2} = 2^{1/2}\varepsilon$. Here, by taking small enough ε, we have that $\hat{v}$ can be arbitrarily close to $\overline{h}(x)$. However we still have $\operatorname{argmax}_{k \in [K]} \left[\overline{h}(x)\right]_k \neq \operatorname{argmax}_{k \in [K]} \left[\hat{v}\right]_k$. Hence, demonstrating proximity under the euclidean norm does not guarantee preservation of the decision made by the argmax.

From the above, we observe that in addition to proximity to the average, the robustness gap induced under robust averaging also depends on point-wise *model dissimilarity* at $x \in \mathcal{X}$, *i.e.*,

$$\sigma_x^2 = \max_{k \in [K]} \frac{1}{n} \sum_{i=1}^{n} \left([h_i(x)]_k - \left[\overline{h}(x)\right]_k\right)^2.$$

We are now ready to present the robustness gap for the case when $\textsc{RobAvg} = \text{CWTM}$. Description of CWTM is deferred to Appendix B.2. The reason for using CWTM is its simplicity and its proven optimality in the formal sense of robust averaging Allouah et al. (2023). To present the result, we introduce some additional notation. For a vector $z \in \Delta^K$, let $z^{(1)}$ and $z^{(2)}$ denote the largest and the 2nd-largest values in the set $\{[z]_k\}_{k=1}^K$. Then, we define $\textsc{margin}(z) = z^{(1)} - z^{(2)}$. If $[z]_k = [z]_{k'}$ for all $k, k' \in [K]$ (i.e., all the coordinates of the vector are equal) then $\textsc{margin}(z) = \infty$.

Theorem 1. *Consider ψ_{rob} as defined in (5) with $\textsc{RobAvg} = \text{CWTM}$. If the regressors $h_1, \dots, h_n$ are such that $\overline{h}(x)$ has a unique maximum coordinate almost everywhere, then the following holds:*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell^{adv}_{\psi_{rob}}(x, \hat{y}_o)\right] \leq \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[\textsc{margin}\left(\overline{h}(x)\right) < 2\left(\sqrt{\frac{\kappa\, n}{n-f}} + \sqrt{\frac{f}{n-f}}\right)\sigma_x\right],$$

where $\hat{y}_o = \psi_o(h_1(x), \dots, h_n(x))$ and $\kappa = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$.

The proof for this theorem is deferred to Appendix B. This shows that the robustness gap for CWTM, when the oracular aggregator is given by the averaging operation, reduces with the fraction of corruptions f/n, the model dissimilarity and the inverse of the average probit's margin. We validate this theoretical finding through an empirical study summarized in Figure 3 (Appendix E.1).

4 DeepSet as oracular aggregator

In this section, we consider the case of more general non-linear trainable (*i.e.*, data dependent) oracular aggregator, inspired from recent work demonstrating the efficacy of such aggregation in context of ensembling (Allouah et al., 2024a; Wang et al., 2024).

Robust empirical federated inference risk minimization (RERM). In this case, since the oracular aggregator ψ_{o} need not be a pre-determined linear combination of input probits (like averaging), we propose to design a robust aggregator ψ_{rob} by directly aiming to minimize the robust federated inference risk $\mathcal{R}_{\text{adv}}(\psi)$. In practice, we seek to minimize the following *robust empirical federated inference risk*:

$$\widehat{\mathcal{R}}_{\text{adv}}(\psi) \coloneqq \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \ell^{\text{adv}}_{\psi}(x, y), \quad \text{where} \quad \ell^{\text{adv}}_{\psi}(x, y) \coloneqq \max_{\mathbf{z}\in\Gamma_f(x)} \mathbb{1}\left\{\psi(\mathbf{z}) \neq y\right\}, \qquad (6)$$

where $\mathcal{D}_{\text{train}}$ comprises of a finite number of i.i.d. data points (x, y) from the global distribution $\mathcal{D}$. In short, we refer to the optimization problem (6) as RERM.

Connection to robustness to adversarial examples. The problem of RERM reduces to the problem of *robustness against adversarial examples* (Madry et al., 2018; Goodfellow et al., 2014), with the input space for the classifier (*i.e.*, aggregator ψ) being a $K \times n$ real-valued matrix with each column in Δ^K and the input perturbation being restricted to corruption of up to f columns of the input matrix. Let $(\Delta^K)^n$ denote the space of such matrices, and for a matrix $M \in (\Delta^K)^n$ let $\delta(M)$ denote the set of matrices from $\mathbb{R}^{K\times n}$ such that $M + V \in (\Delta^K)^n$ for all $V \in \delta(M)$. For any matrix $V \in \mathbb{R}^{K\times n}$, we denote by $\|V\|_0$ the number of non-zero columns (*i.e.*, columns with at least one non-zero entry). We can now formally express the RERM problem in (6) as robustness to adversarial examples as follows:

$$\psi_{\text{rob}} \in \underset{\psi\in\Psi}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \max_{\substack{V\in\delta(H(x)) \\ \|V\|_0\leq f}} \mathbb{1}\left\{\psi\left(H(x)+V\right)\neq y\right\}, \tag{7}$$

where $H(x) = [h_1(x), \ldots, h_n(x)] \in (\Delta^K)^n$. By solving the above adversarial learning problem, we obtain an aggregator with minimum sensitivity against arbitrary perturbation to at most f input probits while ensuring high learning accuracy at the same time. We propose to solve the RERM problem (6) and the equivalent adversarial robustness problem (7) for the space of aggregators Ψ defined by a parameterized deep neural network. However, the adversarial training still remains intractable due to high permutational cardinality of the input space of $H(x)$, totaling to the factor $\sum_{m=1}^{f} {}^nP_m$ since up to any f columns can be perturbed (Liu et al., 2022) where ${}^nP_m = \frac{n!}{(n-m)!}$. To alleviate this issue, we leverage the property that the output of the aggregator must be invariant to the order of columns in $H(x)$. We thus exploit a specific neural network architecture which is permutation invariant to its inputs, as described below.

Robust DeepSet aggregator. Consider $\theta_1 \in \mathbb{R}^{d_1}$ and $\theta_2 \in \mathbb{R}^{d_2}$, and two mappings parameterized by these vectors: $\rho_{\theta_1} : \Delta^K \to \mathbb{R}^p$ and $\mu_{\theta_2} : \mathbb{R}^p \to \Delta^K$. Then, we define Ψ by a set of parameterized mappings $\phi_\theta : (\Delta^K)^n \to \Delta^K$ composed with the argmax operation, where $\theta = (\theta_1, \theta_2)$ and

$$\phi_\theta(\mathbf{z}) \coloneqq \mu_{\theta_2}\left(\frac{1}{n}\sum_{i=1}^{n} \rho_{\theta_1}(z_i)\right). \tag{8}$$

This particular type of neural network is commonly known as DeepSet (Zaheer et al., 2017), in the case when there are no corruptions *i.e.*, $z_i = h_i(x), \forall i \in [n]$. Consequently, in this case, the RERM problem reduces to the following optimization problem:

$$\theta^* \in \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \max_{\mathbf{z}\in\Gamma_f(x)} \mathbb{1}\left\{\underset{k\in[K]}{\operatorname{argmax}} \left[\phi_\theta\left(z_1,\ldots,z_n\right)\right]_k \neq y\right\}. \tag{9}$$

Solving (9) is still complex in practice due to the non-continuity of the indicator function. We address this by changing the point-wise loss function from the indicator function to the cross-entropy loss function, denoted as $\ell(\phi_\theta(\mathbf{z}), y)$. Finally, we still need to compute the maximum value of the point-wise loss at (x, y) over all $\mathbf{z} \in \Gamma_f(x)$. We address this challenge by approximating the maximum value of $\ell(\phi_\theta(\mathbf{z}), y)$ by searching over only a subset of $\mathbf{z} \in \Gamma_f(x)$. Thanks to the permutation invariance property of DeepSet, the permutational cardinality of searching f adversaries now reduces to $\sum_{m=1}^{f}\binom{n}{m}$ where $\binom{n}{m} = \frac{n!}{(n-m)!m!}$. Specifically, in each training iteration, we select $m \leq f$ clients from the set $[n]$ to perturb their probits, repeated N times. We then obtain the perturbed probits and approximate maximum loss by using a multi-step variant of the *fast gradient sign method* (FGSM) for *adversarial training* (Madry et al., 2018). The network parameters are then updated to correctly classify despite the perturbations. In practice, we show that $N \ll \binom{n}{f}$ suffices to obtain a good approximation. Our resulting algorithm is summarized in Algorithm 1.

Finally, we incorporate robust averaging to further reduce the overall sensitivity of ϕ_{θ^*} to probit corruptions. Specifically, with $\theta^* = (\theta_1^*, \theta_2^*)$, the *robust DeepSet aggregator* is defined as follows:

$$\psi_{\text{rob}}(\mathbf{z}) \coloneqq \underset{k\in[K]}{\operatorname{argmax}} \left[\mu_{\theta_2^*}\left(\textsc{RobAvg}(\rho_{\theta_1^*}(z_1),\ldots,\rho_{\theta_1^*}(z_n))\right)\right]_k. \tag{10}$$

Note that we only incorporate robust averaging at the inference time since incorporating at training time renders adversarial training more expensive due to the additional cost of RobAvg.

Algorithm 1: Training of DeepSet Aggregator ϕ_θ

Input: Training Steps E, sampling count N, learning rates γ and η, adversarial sample generation steps S, dataset $\mathcal{D}_{\text{train}}$, number of adversaries f and training loss function ℓ

begin
- **for** 1 **to** E **do**
  - Sample $x, y \sim \mathcal{D}_{\text{train}}$[1]
  - **for** 1 **to** N **do**
    - Choose $m \in \{1, \dots, f\}$ with probability proportional to $\binom{n}{m}$
    - Randomly initialize $\boldsymbol{v} = (v_1, \dots, v_m) \in (\mathbb{R}^K)^m$
    - Sample a random permutation π of $\{1, \dots, n\}$ ▷ *For selecting adversaries*
    - **for** 1 **to** S **do**
      - $\hat{v}_i \leftarrow \text{softmax}(v_i), \forall i \in [m]$ ▷ *Enforce each v_i to be in Δ^K*
      - $\boldsymbol{z} = (h_{\pi(1)}(x), \dots, h_{\pi(n-m)}(x), \hat{v}_1, \dots, \hat{v}_m)$ ▷ *Last m are adversarial*
      - $\boldsymbol{v} \leftarrow \boldsymbol{v} + \gamma \operatorname{sgn}\left(\nabla_{\boldsymbol{v}} \ell\left(\phi_\theta(\boldsymbol{z}, y)\right)\right)$ ▷ *Perturb probits using FGSM*
    - $\hat{v}_i \leftarrow \text{softmax}(v_i), \forall i \in [m]$ ▷ *Final projection of v_i to Δ^K*
    - $\boldsymbol{z} = (h_{\pi(1)}(x), \dots, h_{\pi(n-m)}(x), \hat{v}_1, \dots, \hat{v}_m)$
    - $\theta \leftarrow \theta - \eta \nabla_\theta \ell\left(\phi_\theta(\boldsymbol{z}), y\right)$ ▷ *Approximated maximum loss*
- **return** Trained aggregator ϕ_θ

[1] In practice we sample mini-batches.

Robustness gap of robust DeepSet aggregator. In the theorem below, we present an upper bound on the robustness gap of robust DeepSet aggregator (10), extending Theorem 1 to nonlinear oracular aggregation. In what follows, we make some Lipschitz continuity assumptions. Specifically, there exists real values L_ρ, L_μ such that, for all $z, z' \in \Delta^K$, $v, v' \in \mathbb{R}^p$ and $k \in [K]$,

$$\left|[\rho_{\theta_1^*}(z)]_k - [\rho_{\theta_1^*}(z')]_k\right| \leq L_\rho \left\|z - z'\right\|_2 \ , \ \text{ and}$$
$$\left|[\mu_{\theta_2^*}(v)]_k - [\mu_{\theta_1^*}(v')]_k\right| \leq L_\mu \left\|v - v'\right\|_2 .$$

Let $\mathbf{h}(x) \coloneqq (h_1(x), \dots, h_n(x))$. Recall that $\sigma_x^2 \coloneqq \max_{k \in [K]} \frac{1}{n} \sum_{i=1}^{n} \left([h_i(x)]_k - \left[\overline{h}(x)\right]_k\right)^2$.

Theorem 2. *Consider ψ_{rob} as defined in (10) with* RobAvg = CWTM. *If the regressors $h_1, \dots, h_n$ are such that $\phi_{\theta^*}(\mathbf{h}(x))$ has a unique maximum coordinate almost everywhere, then the following holds:*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell^{adv}_{\psi_{rob}}(x, \hat{y}_o)\right]$$
$$\leq \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[\textsc{margin}\left(\phi_{\theta^*}\left(\mathbf{h}(x)\right)\right) < 2\sqrt{2} L_\mu L_\rho \left(\sqrt{\frac{\kappa n}{n-f}} + \sqrt{\frac{f}{n-f}}\right) \min\left\{1,\ \sqrt{K}\sigma_x\right\}\right],$$

where $\hat{y}_o = \operatorname{argmax}_{k \in [K]} \left[\phi_{\theta^*}\left(\mathbf{h}(x)\right)\right]_k$ and $\kappa = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$.

The proof for this theorem is deferred to Appendix F. Appendix G additionally includes a formal reasoning on the benefits of using RobAvg within DeepSet.

5 Experiments

This section summarizes the key results of our experiments. In all that follows, we consider the RobAvg in the robust DeepSet aggregator (10) to be CWTM, and we refer to the corresponding model as DeepSet-TM. We evaluate DeepSet-TM against robust aggregators like CWMed, against SOTA baselines, and perform a scalability study in Section 5.2. We conduct an ablation study in Appendix E.4, and a comparison with more adversarial defenses in Appendix E.5.

	Aggregation	Logit flipping	SIA-bb	LMA	CPA	SIA	PGD-cw	Worst case
CF-10	Mean	65.4 ± 2.1	59.8 ± 1.1	58.7 ± 3.2	55.6 ± 4.0	42.7 ± 3.9	24.6 ± 4.9	24.6 ± 4.9
	CWMed	56.7 ± 4.3	53.3 ± 2.0	53.8 ± 2.5	52.3 ± 2.9	49.3 ± 3.2	27.8 ± 4.7	27.8 ± 4.7
	GM	63.9 ± 2.3	59.1 ± 1.4	$\mathbf{63.3 \pm 2.3}$	$\mathbf{59.7 \pm 3.2}$	45.3 ± 3.7	25.4 ± 4.8	25.4 ± 4.8
	CWTM	63.3 ± 2.7	59.4 ± 1.3	62.5 ± 2.7	$\mathbf{59.7 \pm 3.2}$	44.8 ± 3.7	27.2 ± 5.1	27.2 ± 5.1
	DeepSet-TM	$\mathbf{67.6 \pm 0.8}$	$\mathbf{62.6 \pm 1.6}$	61.0 ± 4.4	59.4 ± 4.7	$\mathbf{51.4 \pm 2.2}$	$\mathbf{48.2 \pm 4.2}$	48.2 ± 4.2
CF-100	Mean	$\mathbf{78.8 \pm 0.7}$	72.8 ± 1.0	66.6 ± 0.3	66.4 ± 0.3	56.0 ± 1.0	39.2 ± 1.2	39.3 ± 1.3
	CWMed	65.8 ± 1.3	62.3 ± 1.1	66.1 ± 1.2	66.1 ± 1.2	62.9 ± 1.1	41.7 ± 1.3	41.7 ± 1.3
	GM	75.4 ± 0.9	71.5 ± 1.3	75.4 ± 0.8	75.3 ± 0.8	59.6 ± 1.0	39.0 ± 1.3	39.0 ± 1.3
	CWTM	74.8 ± 1.2	71.5 ± 1.3	74.9 ± 1.1	74.8 ± 1.0	60.8 ± 0.9	44.9 ± 1.3	44.9 ± 1.3
	DeepSet-TM	78.0 ± 0.3	$\mathbf{74.7 \pm 0.8}$	$\mathbf{76.0 \pm 0.2}$	$\mathbf{76.4 \pm 0.5}$	$\mathbf{63.7 \pm 0.5}$	$\mathbf{49.6 \pm 0.5}$	49.6 ± 0.5
AG-News	Mean	84.5 ± 0.9	81.4 ± 2.2	$\mathbf{81.2 \pm 2.2}$	76.4 ± 4.0	72.6 ± 4.6	54.9 ± 6.7	54.9 ± 6.7
	CWMed	78.9 ± 3.0	78.4 ± 3.7	75.9 ± 5.0	74.1 ± 4.5	74.4 ± 4.4	53.2 ± 7.0	53.2 ± 7.0
	GM	83.0 ± 0.7	80.7 ± 2.8	80.9 ± 2.5	76.6 ± 3.6	74.0 ± 3.8	52.6 ± 7.3	52.6 ± 7.3
	CWTM	84.3 ± 1.0	81.4 ± 2.2	80.2 ± 2.7	76.4 ± 4.0	72.6 ± 4.6	55.3 ± 7.5	55.3 ± 7.5
	DeepSet-TM	$\mathbf{85.7 \pm 0.4}$	$\mathbf{81.6 \pm 1.6}$	79.2 ± 1.3	$\mathbf{77.5 \pm 1.2}$	$\mathbf{80.1 \pm 1.6}$	$\mathbf{83.2 \pm 1.4}$	77.5 ± 1.2

Table 2: Accuracy (%) of DeepSet-TM against static aggregations on the CIFAR-10, CIFAR-100 and AG-News datasets, with heterogeneity $\alpha = 0.5$, $n = 17$ clients and $f = 4$ adversaries. Logit flipping uses an amplification factor of 2. Results with $\alpha = \{0.3, 0.8\}$ are included in Tables 6 and 7.

5.1 Experimental Setup

Datasets. We evaluate on CIFAR-10, CIFAR-100 (Krizhevsky, 2012), and AG-News (Zhang et al., 2015), covering vision and language tasks. We reserve 10% of training data as server-side validation data for aggregator training, and partition the rest across clients using the Dirichlet distribution $\mathtt{Dir}_n(\alpha)$, in line with previous works (Roux et al., 2025; Dai et al., 2024). Lower α indicates higher heterogeneity. We experiment with $\alpha = \{0.3, 0.5, 0.8\}$. In most of our evaluations, we consider $n = 17$ following prior work (Allouah et al., 2023), except for our scalability study where we vary $n = \{10, 17, 25\}$. Our choice of number of clients is inline with a typical cross-silo federated setting in real-world (Ogier du Terrail et al., 2022).

Models. For CIFAR-10, clients train a ResNet-8 from scratch (He et al., 2016), while for CIFAR-100 and AG-News they fine-tune a ViT-B/32 (Dosovitskiy et al., 2021; Radford et al., 2021) and a DistilBERT (Sanh et al., 2019), respectively. The DeepSet functions μ and ρ are each two-layer MLPs with a ReLU non-linearity. The AutoEncoder in CoPur consists of two-layer MLP encoders and decoders with a leaky ReLU, while the server model is a three-layer MLP (Liu et al., 2022).

Attacks. We evaluate on a total of 6 attacks with varying difficulties depending upon the power of adversary. We consider 4 white-box and 2 black-box attacks where the former assumes access to the server's aggregation model. We propose a new attack called SIA, in the white-box as well as the black-box setting which tries to flip the aggregation decision by exploiting the second most probable class. The remaining attacks constitute the Logit Flipping Attack, Loss Maximization Attack (LMA), Class Prior Attack (CPA) and the Projected Gradient Descent (PGD) attack presented in prior work (Roux et al., 2025; Liu et al., 2022). All attacks and their characteristics are detailed in Appendix D.2, and summarized in Table 4 within. We vary $f \in \{3, 4, 5\}$ across different setups.

Baselines. We compare the performance of DeepSet against several robust aggregators including CWTM (Yin et al., 2018), GM (Pillutla et al., 2022; Small, 1990) and CWMed (Yin et al., 2018) alongside simple averaging. As the data-dependent baselines, we consider CoPur and manifold projection from Liu et al. (2022) where the latter simply projects the input probits onto a learned manifold using an AutoEncoder. Additionally, we also report the performance of non-adversarially trained DeepSet model in our study of robust elements (Appendix E.4).

Reproducibility and reusability. We conduct each experiment with five random seeds, and report the mean and the standard deviation. Our code is also available for reusability[2]. We include additional details on the experimental setup and hyperparameters in Appendix D.

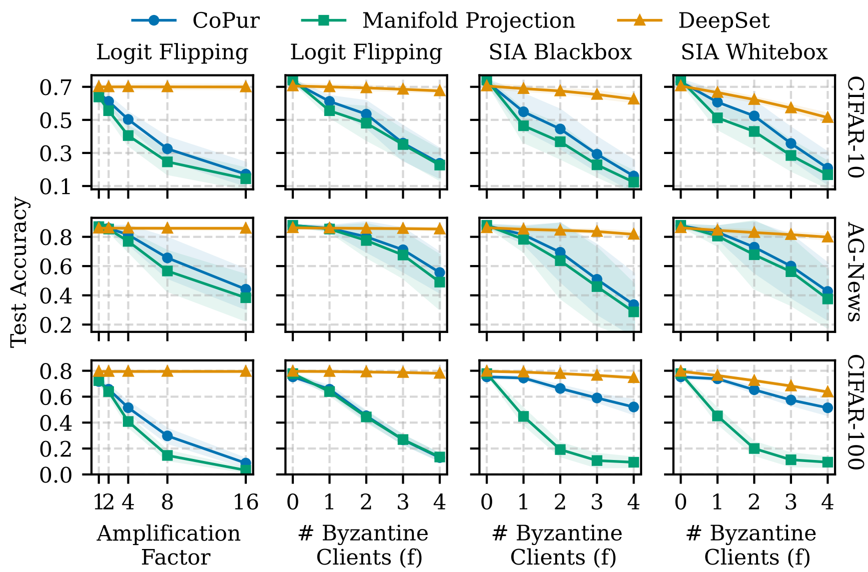

Figure 2: DeepSet-TM vs. baselines under $\alpha = 0.5$ and $n = 17$ clients. In column 1, we have $f = 1$ adversary; other columns use amplification factor 2. See Figure 4 (App. E.6) for results with $\alpha = 0.3$.

n	f	Aggregation	Logit flipping	SIA-bb	LMA	CPA	SIA-wb	PGD-cw	Worst case
10	3	Mean	$\mathbf{75.1 \pm 1.3}$	63.2 ± 2.5	42.5 ± 2.9	42.2 ± 2.8	44.3 ± 1.7	19.7 ± 2.2	19.7 ± 2.2
		CWMed	45.6 ± 2.7	44.5 ± 2.2	52.4 ± 1.6	50.9 ± 1.8	50.5 ± 2.0	16.5 ± 2.3	16.5 ± 2.3
		GM	69.6 ± 1.3	61.3 ± 1.6	$\mathbf{62.6 \pm 2.2}$	$\mathbf{62.2 \pm 2.1}$	47.3 ± 1.5	18.3 ± 2.6	18.3 ± 2.6
		CWTM	69.5 ± 1.3	62.4 ± 2.1	45.2 ± 2.9	44.9 ± 2.8	46.4 ± 1.6	24.0 ± 2.0	24.0 ± 2.0
		DeepSet-TM	73.5 ± 1.3	$\mathbf{66.8 \pm 2.0}$	56.3 ± 2.0	60.0 ± 1.0	$\mathbf{56.5 \pm 1.8}$	$\mathbf{33.2 \pm 2.5}$	33.2 ± 2.5
17	4	Mean	75.6 ± 0.7	69.1 ± 0.8	57.0 ± 2.4	56.8 ± 2.4	44.5 ± 1.4	25.8 ± 2.8	25.8 ± 2.8
		CWMed	55.1 ± 2.1	51.8 ± 2.6	55.9 ± 2.2	55.8 ± 2.2	50.7 ± 1.8	24.5 ± 1.9	24.5 ± 1.9
		GM	71.2 ± 0.8	66.3 ± 1.1	70.2 ± 0.8	70.0 ± 0.9	48.3 ± 1.5	23.8 ± 2.6	23.8 ± 2.6
		CWTM	69.1 ± 1.1	65.8 ± 1.2	69.1 ± 1.2	69.1 ± 1.2	49.6 ± 1.1	32.5 ± 2.4	32.5 ± 2.4
		DeepSet-TM	$\mathbf{75.7 \pm 0.8}$	$\mathbf{71.9 \pm 1.1}$	$\mathbf{72.2 \pm 1.9}$	$\mathbf{72.5 \pm 0.6}$	$\mathbf{56.7 \pm 1.0}$	$\mathbf{38.6 \pm 2.5}$	38.6 ± 2.5
25	5	Mean	$\mathbf{76.0 \pm 1.1}$	71.1 ± 1.3	59.7 ± 2.3	59.4 ± 2.3	40.1 ± 2.3	24.1 ± 2.0	24.1 ± 2.0
		CWMed	51.7 ± 2.9	49.2 ± 2.0	52.1 ± 2.3	52.0 ± 2.3	44.9 ± 1.8	20.6 ± 2.9	20.6 ± 2.9
		GM	70.8 ± 1.4	67.3 ± 1.1	$\mathbf{70.7 \pm 1.5}$	$\mathbf{70.5 \pm 1.5}$	44.2 ± 1.9	21.3 ± 2.0	21.3 ± 2.0
		CWTM	70.1 ± 1.7	67.7 ± 1.4	70.1 ± 1.7	70.1 ± 1.7	45.8 ± 2.0	30.4 ± 2.2	30.4 ± 2.2
		DeepSet-TM	72.8 ± 1.1	$\mathbf{72.8 \pm 1.2}$	65.7 ± 2.3	69.6 ± 1.6	$\mathbf{53.8 \pm 1.5}$	$\mathbf{50.7 \pm 1.5}$	50.7 ± 1.5

Table 3: Scalability study on the CIFAR-100 dataset under $\alpha = 0.3$.

5.2 Results

Performance comparison to Robust Aggregators. Across all datasets in Table 2, DeepSet-TM achieves substantially higher worst-case accuracy (minimum accuracy across attacks) than the robust aggregation baselines. Specifically, it improves over the strongest baseline by +4.7 to +22.2 percentage points depending on the dataset, demonstrating robustness even under the most challenging adversarial conditions. Beyond worst-case performance, DeepSet-TM also shows consistent gains across attacks: in 14 out of 18 dataset–attack combinations, it achieves the highest accuracy, with only small drops ($\leq 2.3\%$ points) in the remaining four cases. The advantage of DeepSet-TM is particularly pronounced against stronger attacks (specifically SIA whitebox and PGD-cw). On CIFAR-10 and CIFAR-100, baseline accuracy drops by 35-40 points, while our approach limits the drop to 20-30. On AG-News, DeepSet-TM retains 83.2% accuracy under PGD-cw compared to 53-55% for the baselines. Notably, as we are working in probit space, giving more power to the adversary (i.e. conducting more PGD iterations during testing than for training) does not lead to performance degradation (see Table 10 in appendix).

[2] https://anonymous.4open.science/r/robust-federated-inference-EC2C

Performance comparison with SOTA baselines. We compare DeepSet-TM with CoPur and manifold projection in Figure 2. Since CoPur operates in logit rather than probit space, it is highly sensitive to input magnitudes, leading to sharp performance degradation under even mild attacks. For instance, in the logit-flipping attack, increasing amplification from 1 to 16 or raising the number of adversarial clients from 1 to 4 reduces CoPur 's accuracy on CIFAR-10 nearly to random chance, while DeepSet-TM remains stable due to its probit-space design and adversarial training. A similar trend holds under the SIA attack, where CoPur suffers large drops on CIFAR-10 and AG-News, though it shows relative robustness on CIFAR-100. This is likely because CoPur leverages block-sparse structure of input probits where higher number of classes induce more sparsity, increasing its effectiveness. Manifold projection performs even worse, as its autoencoder struggles with adversarial patterns spread across many classes. In contrast, DeepSet-TM consistently achieves substantially higher accuracy across all attacks, with only a minor accuracy loss in the no-adversary case ($f = 0$), a known trade-off from adversarial training (Madry et al., 2018).

Scalability study. To assess the generalizability of our approach, we conduct performance evaluations with varying the number of clients: $(n, f) \in \{(10, 3), (17, 4), (25, 5)\}$ on CIFAR-100 under high heterogeneity ($\alpha = 0.3$). In Table 3, we observe that DeepSet-TM consistently provides the highest robustness as the system scales. It improves worst-case accuracy over the best baseline by 9.2%, 6.1% and 20.3% points respectively when increasing n and f. Beyond worst-case performance, DeepSet-TM notably achieves the highest performance in 12 out of 18 dataset–attack combinations. While static aggregations occasionally remain competitive on milder attacks (*e.g.*, LMA), DeepSet-TM is never substantially worse and typically remains close to the highest accuracy across most scenarios.

6 Conclusion

We formalized the problem of robust federated inference and derived a certification for robust averaging under adversarial corruptions. For non-linear aggregation, we proposed DeepSet-TM, a permutation-invariant neural network trained adversarially and combined at test-time with robust averaging. Our experiments demonstrate that DeepSet-TM consistently improves accuracy across datasets and attack types, substantially outperforming prior methods.

REFERENCES

Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4618–4628, 2018.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1232–1300. PMLR, 04 2023. URL https://proceedings.mlr.press/v206/allouah23a.html.

Youssef Allouah, Akash Dhasade, Rachid Guerraoui, Nirupam Gupta, Anne-Marie Kermarrec, Rafael Pinot, Rafael Pires, and Rishi Sharma. Revisiting ensembling in one-shot federated learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=7rWTS2wuYX.

Youssef Allouah, Rachid Guerraoui, Lê-Nguyên Hoang, and Oscar Villemaud. Robust sparse voting. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 991–999. PMLR, 02–04 May 2024b. URL https://proceedings.mlr.press/v238/allouah24a.html.

Xiaoyu Cao, Zaixi Zhang, Jinyuan Jia, and Neil Zhenqiang Gong. Flcert: Provably secure federated learning against poisoning attacks. *Trans. Info. For. Sec.*, 17:3691–3705, January 2022. ISSN 1556-6013. doi: 10.1109/TIFS.2022.3212174. URL https://doi.org/10.1109/TIFS.2022.3212174.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

Tianyue Chu and Nikolaos Laoutaris. Fedqv: Leveraging quadratic voting in federated learning. *SIGMETRICS Perform. Eval. Rev.*, 52(1):91–92, June 2024. ISSN 0163-5999. doi: 10.1145/3673660.3655055. URL https://doi.org/10.1145/3673660.3655055.

Rong Dai, Yonggang Zhang, Ang Li, Tongliang Liu, Xun Yang, and Bo Han. Enhancing one-shot federated learning through data and ensemble co-boosting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tm8s3696Ox.

Arnhav Datar, Arun Rajkumar, and John Augustine. Byzantine spectral ranking. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27745–27756. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1f7288854d3bd476c17725c2d85967f-Paper-Conference.pdf.

Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

Yiqun Diao, Qinbin Li, and Bingsheng He. Towards addressing label skews in one-shot federated learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rzrqh85f4Sc.

Ron Dorfman, Naseem Amin Yehya, and Kfir Yehuda Levy. Dynamic Byzantine-robust learning: Adapting to switching Byzantine workers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11501–11543. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/dorfman24a.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Dongyang Fan, Celestine Mendler-Dünner, and Martin Jaggi. Collaborative learning via prediction consensus. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=SGKbHXoLCI`.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pp. 6246–6283. PMLR, 2022.

Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. Preserving privacy in federated learning with ensemble cross-domain knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11891–11899, 2022.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. *Robust Machine Learning*. Springer, 2024.

Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. *arXiv preprint arXiv:1902.11175*, 2019.

Jenny Hamer, Mehryar Mohri, and Ananda Theertha Suresh. Fedboost: A communication-efficient algorithm for federated learning. In *International Conference on Machine Learning*, pp. 3973–3983. PMLR, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792. URL `https://aclanthology.org/2023.acl-long.792/`.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4585–4593, Apr. 2020. doi: 10.1609/aaai.v34i04.5888. URL `https://ojs.aaai.org/index.php/AAAI/article/view/5888`.

Qinbin Li, Bingsheng He, and Dawn Song. Practical one-shot federated learning for cross-silo setting. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 1484–1490. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/205. URL `https://doi.org/10.24963/ijcai.2021/205`. Main Track.

Wenrui Li, Hanlin Gu, Sheng Wan, Zhirong Luan, Wei Xi, Lixin Fan, Qiang Yang, and Badong Chen. Byzantine robust aggregation in federated distillation with adversaries. In *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, pp. 881–890, 2024. doi: 10.1109/ICDCS60910.2024.00086.

Jing Liu, Chulin Xie, Oluwasanmi O Koyejo, and Bo Li. Copur: Certifiably robust collaborative inference via feature purification. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=r5rzV51GZx.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

May Malka, Erez Farhan, Hai Morgenstern, and Nir Shlezinger. Decentralized low-latency collaborative inference via ensembles on the edge. *IEEE Transactions on Wireless Communications*, 24(1): 598–614, 2025. doi: 10.1109/TWC.2024.3497167.

Darya Melnyk, Yuyi Wang, and Roger Wattenhofer. Byzantine preferential voting. In George Christodoulou and Tobias Harks (eds.), *Web and Internet Economics*, pp. 327–340, Cham, 2018. Springer International Publishing. ISBN 978-3-030-04612-5.

Yuxi Mi, Yutong Mu, Shuigeng Zhou, and Jihong Guan. Fedmdr: Federated model distillation with robust aggregation. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pp. 18–32. Springer, 2021.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, 35:5315–5334, 2022.

Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Christophe Roux, Max Zimmer, and Sebastian Pokutta. On the byzantine-resilience of distillation-based federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=of6EuHT7de.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint*, abs/1910.01108, 2019.

Felix Sattler, Arturo Marban, Roman Rischke, and Wojciech Samek. Cfd: Communication-efficient federated distillation via soft-label quantization and delta coding. *IEEE Transactions on Network Science and Engineering*, 9(4):2025–2038, 2021.

Nir Shlezinger, Erez Farhan, Hai Morgenstern, and Yonina C. Eldar. Collaborative inference via ensembles on the edge. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8478–8482, 2021. doi: 10.1109/ICASSP39728.2021.9414740.

Christopher G Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pp. 263–277, 1990.

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11951–11966, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.698. URL https://aclanthology.org/2024.findings-emnlp.698/.

Hongyi Wang, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric Xing, and Mikhail Yurochkin. Fusing models with complementary expertise. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PhMrGCMIRL.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pp. 5650–5659. Pmlr, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. DENSE: Data-free one-shot federated learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=QFQoxCFYEkA.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 649–657, Cambridge, MA, USA, 2015. MIT Press.