FEDAGREE: LEVERAGING FEDERATED CHECKPOINTS FOR LABEL-FREE OOD EVALUATION VIA AGREEMENT

Giuseppe Serra
Goethe University Frankfurt
German Cancer Consortium (DKTK)*
gserpep@outlook.com

Ben Werner
Goethe University Frankfurt
bwerner.sci@gmail.com

Florian Buettner
Goethe University Frankfurt
German Cancer Consortium (DKTK)*
German Cancer Research Center (DKFZ)
florian.buettner@dkfz-heidelberg.de

ABSTRACT

Federated Learning (FL) has recently emerged as a popular paradigm in many domains, enabling collaborative model training across partners while preserving their privacy. However, distribution shifts in realistic conditions can lead to substantial performance degradation when models are deployed at new sites. Out-of-distribution (OOD) performance estimation is thus critical, but obtaining labelled OOD data is frequently impractical. Let's consider a practical example: in healthcare – where shifts across hospitals are common due to different acquisition devices, patient populations, or clinical protocols – assessing this degradation would be essential, but obtaining labelled data for evaluation is often scarce, time-costly, or too expensive. Agreement-on-the-Line (AotL) (Baek et al., 2022) addresses this by predicting OOD accuracy without labelled data via agreement between pairs of model checkpoints, though obtaining multiple models for this purpose is computationally expensive. We observe that FL naturally resolves this: diverse client checkpoints are already produced during training at no additional cost. We thus propose FEDAGREE, a method to facilitate agreement-based OOD evaluation in federated settings by leveraging both local and cross-client checkpoints. We introduce five checkpoint strategies that progressively expand the use of cross-client information evaluate them across standard OOD benchmarks and diverse medical imaging modalities (dermoscopy, retinopathy, histopathology), under both IID and non-IID settings. Our empirical results demonstrate that FEDAGREE consistently outperforms AotL and confidence-based baselines, confirming that federated settings offer an ideal environment for practical, label-free OOD evaluation.

1 INTRODUCTION

Machine learning models have achieved remarkable success across numerous domains, from image classification on curated benchmarks (He et al., 2016; Dosovitskiy, 2020) to natural language understanding (Devlin et al., 2019). However, a fundamental assumption behind most of these successes is that training and test data are drawn from the same distribution. In practice, this assumption is frequently violated due to, e.g., temporal shifts (Lazaridou et al., 2021), geographical variations (Beery et al., 2020), or changes in data collection procedures (Bandi et al., 2018; Zech et al., 2018). Such distribution shifts can lead to severe performance degradation, undermining the reliability of machine learning models in critical applications.

*partner site Frankfurt, a partnership between DKFZ and UTC Frankfurt-Marburg

This challenge is especially acute in federated learning (FL) (McMahan et al., 2017), where multiple clients collaboratively train models on decentralised data. When these models are later deployed at new client sites, they often encounter data from distributions that differ from those seen during training. In healthcare, for instance, a model trained collaboratively across hospitals may fail at a new site due to differences in acquisition devices, imaging protocols, or patient populations (Daneshjou et al., 2022).

An important question therefore remains open: how can we reliably estimate model performance on out-of-distribution (OOD) data without access to labelled examples from the target distribution? This is particularly relevant since obtaining labels may be expensive, time-consuming or impractical – in healthcare, for example, clinical experts have limited time for accurate annotation of data from new sites.

Agreement-on-the-Line (AotL) (Baek et al., 2022) addresses this challenge by predicting OOD accuracy by measuring agreement between pairs of models trained on similar data distributions, without requiring labelled target data. While the method has demonstrated strong empirical performance across various distribution shift scenarios, it requires multiple trained models, which can be computationally expensive to obtain in sufficient quantity. In this paper, we demonstrate that FL naturally resolves this limitation. The collaborative training process across clients produces diverse model checkpoints at each communication round, and we show that these checkpoints can be leveraged locally for OOD performance estimation at minimal additional cost.

In this paper, we propose FEDAGREE, a method that facilitates agreement-based OOD eestimation in federated settings at minimal cost. Our main contributions are as follows:

- We introduce FEDAGREE, the first method to leverage FL for agreement-based OOD evaluation, exploiting the natural availability of diverse checkpoints across clients.
- We provide a systematic analysis of how local and cross-client checkpoints can be combined for more accurate OOD performance estimation.
- We empirically demonstrate that FEDAGREE outperforms AotL and confidence-based baselines across standard OOD benchmarks and diverse medical imaging modalities in both IID and non-IID settings, confirming that federated settings offer an ideal environment for practical, label-free OOD evaluation.

2 RELATED WORK

2.1 DECENTRALISED FEDERATED LEARNING

Federated Learning (FL) (McMahan et al., 2017) is a distributed learning paradigm where multiple clients collaboratively train a model without sharing their local data. In a standard setting, clients perform local training on their private datasets and periodically communicate model updates to a central server, which aggregates them to produce a global model (centralised FL). While FL research often assumes strict privacy constraints, real-world collaborations frequently involve trusted partners operating under institutional agreements to allow model exchange (decentralised FL (Lalitha et al., 2018; Yuan et al., 2024)).

The decentralised setting is particularly well-suited for applying AotL. Receiving model checkpoints from other clients allows each participant to construct multiple model pairs for agreement computation without the overhead of training additional models locally. This is especially practical in domains such as healthcare, where high-dimensional data is costly to store and replicate. By transferring models rather than data, FL scales naturally with growing datasets without disproportionate storage overhead (Rieke et al., 2020).

In this work, under this practical setting, we explore how clients can effectively leverage both local and external checkpoints to improve OOD accuracy estimation. We evaluate both IID scenarios, with data uniformly distributed across clients, and non-IID settings, where each client has data from distinct domains or sources, reflecting federated scenarios with heterogeneous data distributions.

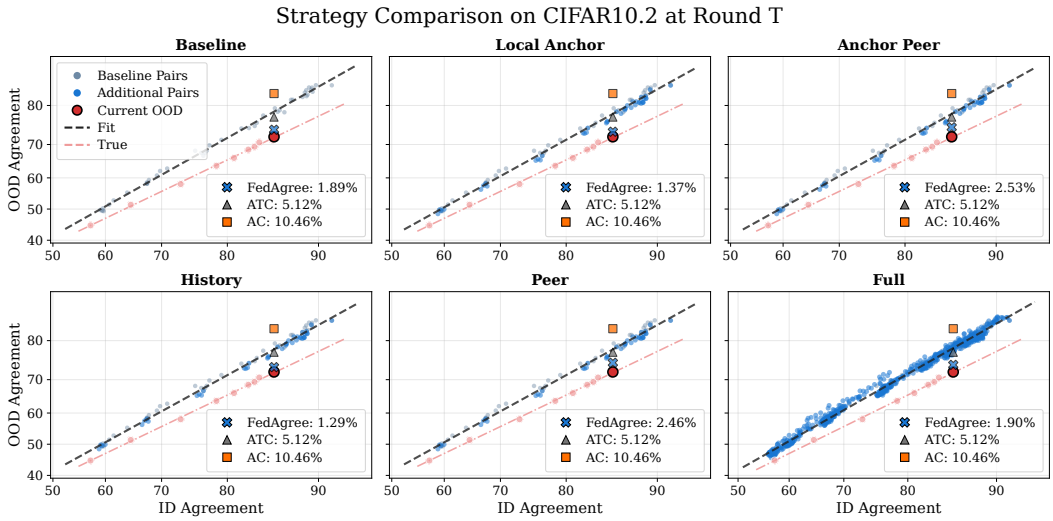

Figure 1: Visual comparison of the FEDAGREE checkpoint strategies on a local client at round T. Each strategy progressively increases the number of agreement points used to fit the AotL relationship, from only local checkpoints (Baseline) to all available checkpoints across all clients and rounds (Full). All strategies show strong linear correlation between ID and OOD agreement, with more agreement points generally yielding more robust estimates (×) which are closer to the true OOD accuracy (●) compared to confidence-based methods (ATC (▲) and AC (■)).

2.2 OOD GENERALISATION

OOD generalisation refers to the ability of models to maintain performance when deployed on data differing from the training distribution (Liu et al., 2021; Zhou et al., 2022). Domain generalisation methods (Muandet et al., 2013; Li et al., 2017; Gulrajani & Lopez-Paz, 2021) aim to learn domain-invariant representations, while domain adaptation techniques (Ganin et al., 2016; Long et al., 2018; Hoyer et al., 2023) leverage unlabelled target data to adapt models at deployment. Test-time adaptation (Wang et al., 2021; 2022; Niu et al., 2022) further allows adaptation using only test data.

A complementary line of research focuses on estimating model performance under distribution shift without labelled target data, including methods based on confidence calibration (Guo et al., 2017; Hendrycks & Gimpel, 2017) and density estimation (Morteza & Li, 2022; Peng et al., 2024; Koebler et al., 2025). Baek et al. (2022) introduced Agreement-on-the-Line (AotL), which predicts OOD accuracy by measuring agreement between model pairs on unlabelled target data, exploiting the linear relationship between source-distribution agreement and target accuracy. AotL assumes a centralised setting where multiple models are available. In contrast, FEDAGREE leverages the naturally occurring model diversity in FL, reducing computational burden while potentially improving estimation accuracy through access to models trained on non-overlapping data subsets.

3 METHODOLOGY

3.1 PROBLEM SETUP AND NOTATION

We consider a decentralised federated setting with K clients, each holding a local dataset $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{n_k}$ drawn from a shared source distribution P_S. At each communication round t, clients perform local training and exchange model checkpoints. We denote client k's parameters at round t as $\theta_k^{(t)}$. After T rounds, each client has access to its own checkpoints $\{\theta_k^{(t)}\}_{t=1}^{T}$ and those received from other clients $\{\theta_j^{(t)}\}_{j\neq k, t=1}^{T}$.

At deployment, each client may encounter data from a target distribution $P_T \neq P_S$. Our goal is to estimate OOD accuracy without labelled target data. In practice, distribution shifts arise natu-

rally across clients – for example, different hospitals using distinct imaging equipment, or different geographic regions with varying data collection procedures. Since each client has labelled ID validation data, $\text{Acc}_{\text{ID}}(\theta)$ can be computed directly, while $\text{Acc}_{\text{OOD}}(\theta)$ requires labels from P_T, which are assumed to be unavailable.

Agreement-on-the-Line (AotL) (Baek et al., 2022) addresses this by replacing accuracy with *agreement* between model pairs – a label-free metric. Given two models θ_i and θ_j, their agreement on dataset $\mathcal{D}$ is:

$$\text{Agr}(\theta_i, \theta_j) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{1}[f_{\theta_i}(x) = f_{\theta_j}(x)]. \tag{1}$$

Agr_{ID} is evaluated on validation inputs (ignoring labels); Agr_{OOD} requires only unlabelled samples from P_T.

Agreement-on-the-Line. Baek et al. (2022) observed that whenever probit-scaled[1] ID and OOD accuracy exhibit a strong linear correlation across models (accuracy-on-the-line (Miller et al., 2021)), the probit-scaled ID and OOD *agreement* also exhibits a strong linear correlation with approximately matching slope and bias:

$$\begin{aligned} &\Phi^{-1}(\text{Acc}_{\text{OOD}}(\theta_i)) \approx a \cdot \Phi^{-1}(\text{Acc}_{\text{ID}}(\theta_i)) + b \iff \\ &\Phi^{-1}(\text{Agr}_{\text{OOD}}(\theta_i, \theta_j)) \approx a \cdot \Phi^{-1}(\text{Agr}_{\text{ID}}(\theta_i, \theta_j)) + b \end{aligned} \tag{2}$$

When accuracy-on-the-line does not hold, agreement-on-the-line also fails, providing a built-in reliability check.

Regression fitting. Given M models $\{\theta_m\}_{m=1}^M$, slope $\hat{a}$ and bias $\hat{b}$ are estimated via linear regression on probit-scaled pairwise agreements:

$$\hat{a}, \hat{b} = \arg\min_{a,b \in \mathbb{R}} \sum_{i \neq j} \left(\Phi^{-1}(\text{Agr}_{\text{OOD}}(\theta_i, \theta_j)) - a \cdot \Phi^{-1}(\text{Agr}_{\text{ID}}(\theta_i, \theta_j)) - b \right)^2 \tag{3}$$

OOD accuracy prediction. Since agreement and accuracy lines share approximately the same slope and bias, OOD accuracy is predicted as:

$$\widehat{\text{Acc}}_{\text{OOD}}(\theta_m) = \Phi\left(\hat{a} \cdot \Phi^{-1}(\text{Acc}_{\text{ID}}(\theta_m)) + \hat{b} \right) \tag{4}$$

In our federated setting, having access to peer checkpoints, the prediction can be further regularised by replacing $\Phi^{-1}(\text{Acc}_{\text{ID}}(\theta_m))$ in Eq. 4 with the mean over the target and reference models' ID accuracies, to account for potentially unreliable local estimates. Thus, the updated equation for OOD accuracy prediction reads:

$$\widehat{\text{Acc}}_{\text{OOD}}(\theta_m) = \Phi\left(\hat{a} \cdot \left(\frac{1}{M} \sum_{m=1}^{M} \left(\Phi^{-1}(\text{Acc}_{\text{ID}}(\theta_m)) \right) \right) + \hat{b} \right) \tag{5}$$

3.2 FEDAGREE

AotL requires a set of models to compute pairwise agreements for both fitting the linear relationship (Eq. 3) and predicting OOD accuracy (Eq. 4). In a centralised setting, this typically requires training multiple models from scratch, which is computationally expensive. We find that decentralised FL naturally alleviates this cost: after T rounds, client k has access to a rich model pool comprising its own history $\{\theta_k^{(t)}\}_{t=1}^T$ and all received checkpoints $\{\theta_j^{(t)}\}_{j \neq k, t=1}^T$, *without any additional training overhead*, making it practically useful.

The available checkpoints can be used at two stages of the AotL pipeline: *regression*, where pairwise agreements fit the linear relationship (Eq. 3), and *prediction*, where the fitted line estimates OOD accuracy (Eq. 5). We investigate six strategies that progressively expand the use of federated checkpoints. Incorporating cross-client checkpoints is motivated by two factors: the quality of the linear fit improves with more diverse model pairs, and models from other clients introduce additional diversity from different local data and training trajectories. We describe each strategy from the perspective of client k at round T.

[1]The probit function $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution, applied to induce a better linear fit (Miller et al., 2021).

Table 1: Summary of checkpoint strategies for client k at round T with K clients. *Regression* refers to which model pairs are used to fit the agreement line; *Prediction* refers to which models are compared against $\theta_k^{(T)}$ for OOD accuracy estimation.

Strategy	Regression pairs	Prediction comparisons	# Regression	# Prediction
AotL	Own history only	Own history	$\binom{T}{2}$	$T-1$
Local Anchor	History + Hist.↔Current	Own history	$\binom{T-1}{2}+(T-1)K$	$T-1$
Anchor Peer	History + Hist.↔Current	History + Current peers	$\binom{T-1}{2}+(T-1)K$	$T+K-2$
History	All pairs in pool	Own history	$\binom{T+K-1}{2}$	$T-1$
Peer	All pairs in pool	History + Current peers	$\binom{T+K-1}{2}$	$T+K-2$
Full	All pairs (global)	All other checkpoints	$\binom{KT}{2}$	$KT-1$

AotL – Local. Uses only the client's own checkpoints $\{\theta_k^{(t)}\}_{t=1}^{T}$, without any cross-client information. This serves as a reference for what any client can achieve independently and reflects the original idea presented in Baek et al. (2022).

Local Anchor. Exploits cross-client information for regression while keeping prediction local. The pool includes the client's own history and current-round checkpoints from other clients $\{\theta_j^{(T)}\}_{j\neq k}$. Regression uses all historical pairs and history-to-current pairs, but excludes current-to-current cross-client pairs, as these models – having just completed the same round – may be too similar to provide informative agreement variation. Prediction remains local.

Anchor Peer. Shares the same regression procedure as Local Anchor but extends prediction to include cross-client models: $\theta_k^{(T)}$ is compared against both its own history $\{\theta_k^{(t)}\}_{t=1}^{T-1}$ and current-round checkpoints from other clients $\{\theta_j^{(T)}\}_{j\neq k}$.

History. Same pool and regression setup as Local Anchor, but removes the restriction on which pairs are used for regression, including all pairwise combinations (including current-to-current cross-client pairs). Prediction remains local.

Peer. Extends both regression and prediction to use cross-client models. Regression uses all pairwise combinations; prediction compares $\theta_k^{(T)}$ against both its own history and current-round peer checkpoints.

Full. All checkpoints $\{\theta_j^{(t)}\}_{j=1,\ldots,K,\ t=1,\ldots,T}$ are pooled, making maximal use of the federated setting. All pairwise agreements are used for regression, and prediction compares $\theta_k^{(T)}$ against all other checkpoints, yielding $\binom{KT}{2}$ regression pairs.

Table 1 summarises the six strategies, forming a progression from fully local evaluation (AotL) to full federated exploitation (Full). By comparing them, we isolate the marginal benefit of cross-client checkpoints at each stage of the AotL pipeline.

4 Experiments

4.1 Datasets

We evaluate FedAgree on OOD benchmarks spanning different types of distribution shift, ranging from standard computer vision testbeds to clinically relevant medical imaging scenarios.

CIFAR-10.1 (Recht et al., 2018) and **CIFAR-10.2** (Lu et al., 2020) are reproductions of the CIFAR-10 (Krizhevsky et al., 2009) test set intended to measure how well models generalise beyond the original test distribution. Models are trained on standard CIFAR-10 (ID) and evaluated on each vari-

ant separately (OOD). These benchmarks represent relatively mild, naturally occurring distribution shifts and are standard testbeds for AotL (Baek et al., 2022).

We then evaluate on three medical imaging benchmarks that reflect realistic federated deployment scenarios, each with a distinct clinically relevant distribution shift. **HAM10000** (ID) (Tschandl et al., 2018; Yang et al., 2023) and **BCN20000** (OOD) (Hernández-Pérez et al., 2024) are dermoscopic imaging datasets for skin lesion classification, where the shift arises from differences in acquisition devices and patient populations across dermatology centres. **DeepDRiD** (ID) (Liu et al., 2022; Yang et al., 2023) and **APTOS** (OOD) (Karthik et al., 2019) are fundus photography datasets for diabetic retinopathy grading, where the shift stems from differences in imaging protocols across clinical sites.

Finally, for non-IID experiments, we use **Camelyon17-WILDS** (Bandi et al., 2018; Koh et al., 2021), a histopathological dataset for tumour detection across five hospitals. This is a natural federated scenario where each client corresponds to a different hospital, with distribution shift arising from differences in staining and imaging equipment – making it an ideal testbed for the non-IID setting.

4.2 Experimental Setup

Federated setup and training. We simulate a decentralised federated setting with $K = 4$ clients and $T = 10$ communication rounds. For IID experiments (Table 2), the ID training data is partitioned uniformly at random across clients. For the medical imaging datasets (HAM10000, DeepDRiD, Camelyon17), we re-sample the OOD datasets to approximatively match the class proportions of their respective ID counterparts, isolating the distribution shift of interest from confounding label shift. For non-IID experiments on Camelyon17 (Table 3), each client corresponds to a different hospital and is evaluated on the held-out hospital – reflecting the realistic federated scenario where heterogeneous data naturally arises from different institutions. Clients exchange checkpoints at each round and retain all received checkpoints for agreement-based evaluation. We use FedAvg (McMahan et al., 2017) as the aggregation strategy, SlimResNet18 for CIFAR-10, and ResNet18 (He et al., 2016) for all medical imaging datasets. All models are trained with batch size 128 (CIFAR-10, Camelyon17) or 32 (HAM10000, APTOS) and learning rate 0.1 (CIFAR-10), 0.01 (APTOS, Camelyon17), or 0.001 (HAM10000). Code is available at: `https://github.com/MLO-lab/FedAgree`.

Evaluation metric. Following Baek et al. (2022), we report the Mean Absolute Error (MAE) between the predicted and true OOD accuracy, expressed in percentage points. Lower MAE indicates more accurate OOD performance estimation. For each experiment, we compute the MAE at the last communication round, average across all clients, and report the mean and standard deviation over three random seeds.

Baselines. We compare the five FedAgree strategies against Average Confidence (AC) (Hendrycks & Gimpel, 2017), which uses the mean softmax probability as a proxy for accuracy, and Average Threshold Confidence (ATC) (Garg et al., 2022), which thresholds a confidence score to estimate the fraction of correct predictions. Both baselines operate on individual models and require no additional checkpoints. We also include the original AotL baseline, which applies ALine-D using only each client's own historical checkpoints, to verify whether incorporating cross-client information provides a benefit over local-only agreement-based estimation.

4.3 Experimental Results

Table 2 presents IID results across four benchmarks. FedAgree consistently outperforms both confidence-based baselines (AC, ATC) and the local-only AotL baseline across all settings.

Cross-client checkpoints consistently improve over AotL. All FedAgree strategy outperforms AotL in every benchmark, with the Full strategy providing the most consistent gains. The benefit is particularly pronounced on distribution shifts reflecting realistic clinical deployment challenges: on BCN20000 (dermoscopy), Full achieves 44% MAE reduction over AotL and 67% over AC; on APTOS (retinopathy), Full reduces MAE by 52% over AotL and 85% over AC. These large gains on medical imaging datasets illustrate a practically important scenario: when federated clients must

Table 2: Mean Absolute Error (MAE ↓) of OOD accuracy estimation across four benchmarks. We compare single-model baselines (AC, ATC) and the local-only baseline (AotL) against five FEDAGREE checkpoint strategies. Best performance is **highlighted**, second best is underlined.

		Mean Absolute Error (MAE ↓)			
Method	**Checkpoint Strategy**	**CIFAR10.1**	**CIFAR10.2**	**BCN20000**	**APTOS**
AC	—	7.21 ± 0.13	10.22 ± 0.30	16.73 ± 1.33	39.98 ± 0.98
ATC	—	1.52 ± 0.13	4.28 ± 0.35	9.05 ± 3.05	8.54 ± 0.94
AotL	—	2.99 ± 0.08	6.05 ± 0.25	5.29 ± 1.08	12.18 ± 1.56
FEDAGREE	Local Anchor	2.30 ± 0.11	0.87 ± 0.30	3.80 ± 0.31	8.77 ± 1.68
	Anchor Peer	**1.05 ± 0.12**	2.10 ± 0.29	3.92 ± 0.31	8.99 ± 1.47
	History	2.36 ± 0.08	**0.78 ± 0.27**	3.80 ± 0.23	8.77 ± 1.63
	Peer	1.11 ± 0.06	2.01 ± 0.26	3.96 ± 0.33	8.97 ± 1.44
	Full	1.80 ± 0.13	1.30 ± 0.28	**2.96 ± 0.90**	**5.85 ± 1.64**

Table 3: Mean Absolute Error (MAE ↓) for Non-IID experiments on Camelyon17-WILDS. Results report the MAE at the last communication round, averaged across clients and three random seeds. Best performance is **highlighted**, second best is underlined.

		Mean Absolute Error (MAE ↓)
Method	**Checkpoint Strategy**	**Camelyon (Non-IID)**
AC	—	31.45 ± 4.59
ATC	—	32.59 ± 2.88
AotL	—	34.67 ± 3.30
FEDAGREE	Local Anchor	17.43 ± 2.04
	Anchor Peer	16.02 ± 2.11
	History	17.84 ± 1.93
	Peer	16.44 ± 2.01
	Full	**14.80 ± 3.75**

generalise across different acquisition devices, patient populations, or clinical sites, the checkpoint diversity introduced by FL is particularly valuable for reliable OOD estimation.

Agreement-based methods outperform confidence-based baselines. Single-model confidence baselines (AC and ATC) are consistently outperformed by agreement-based approaches across all benchmarks. The advantage is particularly pronounced on challenging distribution shifts – on CIFAR-10 under mild shift, ATC remains competitive, but on the more severe shifts present in medical imaging, agreement-based methods offer substantially lower error. These results confirm that pairwise agreement provides a stronger signal for OOD accuracy estimation than confidence-based methods, and that federated settings – where multiple checkpoints are naturally available – are particularly well-suited for this approach.

Strategy selection depends on distribution shift characteristics. Under mild distribution shifts (CIFAR-10 results), gains are smaller and selective strategies (Anchor Peer, History) achieve better performance. While no single strategy dominates in all settings, all FEDAGREE variants improve over AotL, suggesting that practitioners can adopt FEDAGREE without extensive hyperparameter tuning, with Full serving as a reliable default choice, especially under severe or unknown distribution shifts.

The Non-IID Case. Table 3 reports non-IID results on Camelyon17-WILDS, where each client represents a different hospital. This reflects a broader class of federated scenarios where heterogeneous client data is the norm rather than the exception – arising in any setting where clients correspond to distinct institutions, geographies, or data collection pipelines. In this realistic setting, the benefits of cross-client checkpoints become even more pronounced: all FEDAGREE strategies substantially outperform AC, ATC, and AotL. Heterogeneous client data yields checkpoints that capture com-

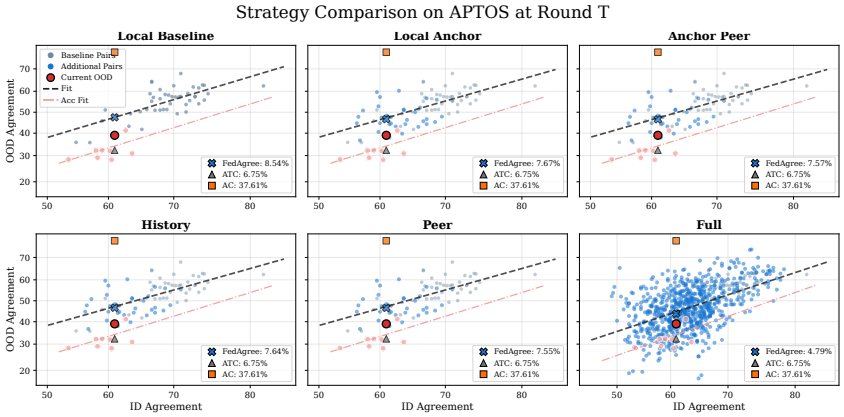

Figure 2: Visual comparison of the FEDAGREE checkpoint strategies on a local client at round T on the APTOS dataset. Unlike CIFAR10.2, the linear relationship between ID and OOD agreement is less pronounced, reflecting the greater distributional shift inherent to the dataset. Nevertheless, incorporating additional checkpoints progressively refines the fit of the AotL relationship, yielding more robust estimates (×) that better approximate the true OOD accuracy (●) compared to confidence-based methods (ATC (▲) and AC (■)).

plementary model behaviours, providing the diversity needed for more precise OOD estimation. FEDAGREE consistently outperforms all baselines, despite lower R^2 values (0.45–0.72) indicating weaker linear fits that suggest estimates should be treated with caution.

5 DISCUSSION AND CONCLUSION

In this paper, we proposed FEDAGREE, a method that leverages the natural availability of model checkpoints in decentralised federated learning for label-free OOD accuracy estimation via Agreement-on-the-Line (AotL). The core observation is general: any federated system that exchanges model checkpoints during training already possesses the diverse model pool required for agreement-based evaluation, at no additional cost. We introduced five checkpoint strategies (Table 1) that progressively expand the use of cross-client information across the regression and prediction stages of the AotL pipeline, and evaluated them across diverse OOD scenarios under both IID and non-IID data partitions (Tables 2 and 3).

Our results demonstrate that incorporating cross-client checkpoints consistently improves OOD accuracy estimation over purely local evaluation (AotL), and that agreement-based methods considerably outperform confidence-based baselines (AC, ATC) across all benchmarks. The gains are especially pronounced in the medical imaging experiments, where distribution shifts reflect realistic deployment challenges such as differences in acquisition devices, imaging protocols, and patient populations across clinical sites – a setting where label-free evaluation is particularly valuable. We further investigated the coefficient of determination R^2 as a practical sanity check for assessing the reliability of OOD estimates (Tables 4 and 5): higher R^2 values indicate more trustworthy estimates, while low R^2 signals that practitioners should interpret results with caution – though FEDAGREE still outperforms available alternatives even in such cases (Figure 2).

Our findings suggest that the optimal strategy depends on the nature of the distribution shift. Selective strategies work well under mild shifts where the linear assumption holds strongly, while Full – which maximises checkpoint diversity – performs best under severe distribution shifts and is *always* better than local-only solutions. We therefore recommend Full as a safe default in any federated deployment scenario.

Limitations and future work. While we investigated both IID and non-IID settings, our experiments assume a fixed federated strategy. Exploring how FEDAGREE interacts with other federated components (e.g., aggregation strategies, local training schedules, and privacy-preserving mechanisms) would further clarify its practical applicability across a wider range of domains.

ACKNOWLEDGMENTS

This work was supported by the The Federal Ministry for Economic Affairs and Climate Action of Germany (BMWK, Project OpenFLAAS 01MD23001E). Co-funded by the European Union (ERC, TAIPO, 101088594). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31): eabq6147, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific data*, 11(1):641, 2024.

Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11721–11732, 2023.

Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. https://kaggle.com/competitions/aptos2019-blindness-detection, 2019. Kaggle.

Alexander Koebler, Thomas Decker, Ingo Thon, Volker Tresp, and Florian Buettner. Incremental uncertainty-aware performance monitoring with active labeling intervention. In *International Conference on Artificial Intelligence and Statistics*, pp. 2188–2196. PMLR, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

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

Anusha Lalitha, Shubhanshu Shekhar, Tara Javidi, and Farinaz Koushanfar. Fully decentralized federated learning. In *Third workshop on bayesian deep learning (NeurIPS)*, volume 12, 2018.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363, 2021.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6), 2022.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, volume 5, pp. 15, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pp. 7721–7735. PMLR, 2021.

Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7831–7840, 2022.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pp. 10–18. PMLR, 2013.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. Conjnorm: Tractable density estimation for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 10(1):41, 2023.

Liangqi Yuan, Ziran Wang, Lichao Sun, Philip S Yu, and Christopher G Brinton. Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*, 11(21):34617–34638, 2024.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

A APPENDIX

A.1 AGREEMENT FIT QUALITY WITH R^2

Agreement fit quality as a reliability indicator. Baek et al. (2022) observed that AotL performs well when the R^2 of the agreement fit is high (≥ 0.95) and degrades when the linear relationship is weak ($R^2 < 0.75$), but note that competing baselines perform even worse in such cases. We observe the same pattern in our federated setting (Tables 4 and 5). On CIFAR-10.1 and CIFAR-10.2, where $R^2 > 0.99$, all FEDAGREE strategies achieve low MAE ($\leq 2.30\%$). On the medical imaging benchmarks (BCN20000, APTOS) and Camelyon17-WILDS, where R^2 drops substantially, the estimation error increases. However, even in these challenging cases, FEDAGREE consistently outperforms both AC and ATC. This suggests that R^2 can serve as a practical sanity check: when the agreement fit is strong, practitioners can trust the OOD estimate with high confidence; when it is weak, the estimate should be treated with caution, though it remains preferable to available alternatives.

Table 4: Coefficient of determination (R^2) of the agreement line fit for IID benchmarks. Higher values indicate a stronger linear relationship between ID and OOD agreement. Strategies with the same regression pool yield identical R^2 values. Results averaged across clients and random seeds.

		R^2 **of Agreement Fit** ↑	
Method	**Checkpoint Strategy**	**CIFAR10.1**	**CIFAR10.2**
AotL	—	99.37 ± 0.11	99.57 ± 0.04
FEDAGREE	Local Anchor / Anchor Peer	99.21 ± 0.10	99.39 ± 0.03
	History / Peer	99.19 ± 0.10	99.37 ± 0.06
	Full	99.06 ± 0.13	99.29 ± 0.01

Table 5: Coefficient of determination (R^2) of the agreement line fit for Camelyon17-WILDS (Non-IID). Results averaged across clients and random seeds.

		R^2 **of Agreement Fit** ↑
Method	**Checkpoint Strategy**	**Camelyon (Non-IID)**
AotL	—	51.45 ± 3.06
FEDAGREE	Local Anchor / Anchor Peer	71.93 ± 5.05
	History / Peer	69.62 ± 6.56
	Full	46.62 ± 4.98