FedAgree: Label-Free Performance Estimation under Distribution Shift for Federated Medical Imaging Analysis

Giuseppe Serra[1,2,3]✉, Ben Werner[3], and Florian Buettner[1,2,3]

[1] German Cancer Research Center (DKFZ), Heidelberg, Germany
name.surname@dkfz-heidelberg.de
[2] German Cancer Consortium (DKTK), partner site Frankfurt, a partnership between DKFZ and UTC Frankfurt-Marburg
[3] Goethe University, Frankfurt, Germany

Abstract. Federated Learning (FL) has recently emerged as a popular paradigm in healthcare, enabling collaborative model training across hospitals while preserving patient privacy. However, distribution shifts arising from different acquisition devices, patient populations, or clinical protocols can lead to substantial performance degradation when models are deployed at new sites. Out-of-distribution (OOD) performance estimation is thus critical to assess this degradation, but labelled OOD data from target hospitals is scarce or expensive to obtain. Agreement-on-the-Line (AotL) [1] 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 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 and evaluate them on diverse medical imaging modalities (dermoscopy, retinopathy, histopathology), under both IID and non-IID settings. Our empirical results demonstrate that federated healthcare settings offer an ideal environment for practical, label-free OOD evaluation.

Keywords: OOD Generalisation · Medical Federated Learning

1 Introduction

Federated Learning (FL) [21] has become an essential paradigm for collaborative healthcare AI, enabling hospitals to jointly train models while preserving patient privacy. However, these models often face clinically relevant distribution shifts when deployed at new sites due to differences in acquisition devices (e.g., scanner manufacturers, camera models), imaging protocols (e.g., patient positioning, tissue staining procedures), or patient populations (e.g., demographics).

For instance, a melanoma detection model may misclassify malignant lesions at a new clinic simply because lighting conditions differ from its training environment [3]. Thus, such distribution shifts can lead to severe performance degradation, undermining the reliability of AI systems in critical healthcare applications.

Out-of-distribution (OOD) generalisation is therefore crucial to assess model performance before clinical deployment. However, obtaining labelled data from target hospitals is often expensive, time-consuming or even impractical since clinical experts have limited time for accurate label annotation. Thus, an important question remains open: how can we reliably estimate model performance on OOD data during deployment without labelled examples from the target distribution?

Agreement-on-the-Line (AotL) [1] 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. This has hindered adoption in real-world clinical applications. In this paper we demonstrate that FL can naturally address this limitation. Collaborative training process across different clinical sites produces diverse model checkpoints at each communication round and we demonstrate 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 performance estimation in federated settings at minimal cost. Our main contributions are:

- We introduce FedAgree, the first method for agreement-based OOD performance estimation in federated settings, exploiting the natural availability of diverse checkpoints across different clinical sites.
- 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 agreement-on-the-Line (AotL) and confidence-based baselines on diverse medical imaging modalities in both IID and non-IID settings.

2 Related Work

2.1 Decentralised Federated Learning

Federated Learning (FL) [21] is a distributed learning paradigm where multiple clients collaboratively train a model without sharing their local data. While FL research often assumes strict privacy constraints, real-world healthcare collaborations frequently involve trusted partners operating under institutional agreements to allow model exchange (decentralised FL [15]). This is particularly advantageous in healthcare, where high-dimensional medical imaging data is

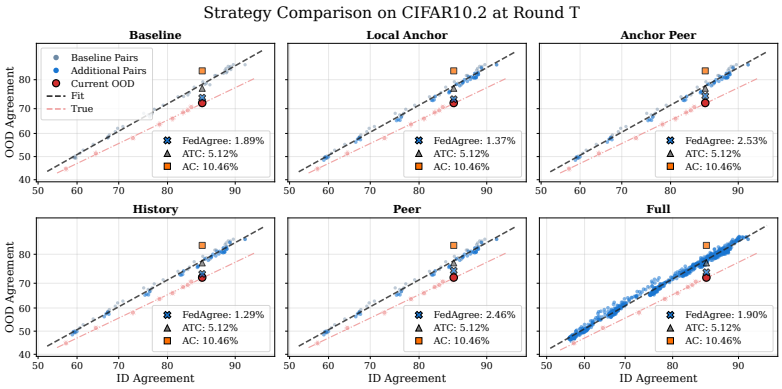

Fig. 1. Visual comparison of the FedAgree checkpoint strategies on a local client at round T. Each strategy progressively increases the number of agreement points used to fit the AotL relationship, from only local checkpoints (Baseline) to all available checkpoints across all clients and rounds (Full). 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 (■)).

costly to store and replicate. By transferring models rather than data, FL scales naturally with growing datasets without disproportionate storage overhead [27].

Under this practical setting, we explore how clients can leverage both local and external checkpoints to improve OOD performance estimation. We evaluate both IID and non-IID scenarios, where each client has data from distinct domains, reflecting federated settings with heterogeneous data distributions.

2.2 OOD Generalisation

OOD generalisation refers to the ability of models to maintain performance when deployed on data differing from the training distribution [17,30]. Domain generalisation methods [24,16,6] aim to learn domain-invariant representations, while domain adaptation techniques [4,19,11] leverage unlabelled target data to adapt models at deployment. A complementary line of research focuses on estimating model performance under distribution shift without labelled target data, including methods based on confidence calibration [7,9], alongside density-based approaches developed for OOD detection [23,25,13]. Baek et al. [1] 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. Since each client has labelled ID validation data, $\mathrm{Acc}_{\mathrm{ID}}(\theta)$ can be computed directly, while $\mathrm{Acc}_{\mathrm{OOD}}(\theta)$ requires labels from P_T, which are unavailable.

Agreement-on-the-Line (AotL) [1] 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:

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

$\mathrm{Agr}_{\mathrm{ID}}$ is evaluated on validation inputs (ignoring labels); $\mathrm{Agr}_{\mathrm{OOD}}$ requires only unlabelled samples from P_T.

Agreement-on-the-Line. Baek et al. [1] observed that whenever probit-scaled[4] ID and OOD accuracy exhibit a strong linear correlation across models (accuracy-on-the-line [22]), the probit-scaled ID and OOD *agreement* also exhibits a strong linear correlation with approximately matching slope and bias:

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

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:

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

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

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

[4] The probit function $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution, applied to induce a better linear fit [22].

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

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	$\binom{KT}{2}$	$KT-1$

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

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. (4)). All strategies use Eqs. (3) and (4) and differ in which checkpoints are supplied to each stage.

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.

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. [1].

FedAgree - 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.

FedAgree - 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 and current-round checkpoints from other clients.

FedAgree - History. Same as Local Anchor but removes the restriction on regression pairs, including all pairwise combinations. Prediction remains local.

FedAgree - 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.

FedAgree - Full. All checkpoints $\{\theta_j^{(t)}\}_{j=1,\dots,K,\ t=1,\dots,T}$ are pooled, with 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 AoTL (i.e., local models only) to full federated exploitation (i.e., FedAgree - Full). By comparing them, we isolate the marginal benefit of cross-client checkpoints at both stages of agreement-based performance estimation.

4 Experiments

4.1 Datasets

To understand the benefit of using FedAgree, we first evaluate it on **CIFAR-10.1** (OOD) [26] and **CIFAR-10.2** (OOD) [20], standard AotL testbeds [1] with mild distribution shifts of **CIFAR10** (ID). We then evaluate on diverse medical imaging modalities reflecting realistic clinical scenarios with three distinct clinically relevant distribution shifts from (i) different acquisition devices, (ii) imaging protocols, and (iii) patient populations. **HAM10000** (ID) [28,29] and **BCN20000** (OOD) [10] are dermoscopic imaging datasets for skin lesion classification, **DeepDRiD** (ID) [18,29] and **APTOS** (OOD) [12] are fundus photography datasets for diabetic retinopathy grading. Finally, we consider non-IID splits on one additional benchmark: **Camelyon17-WILDS** [2,14] a histopathological dataset for tumor detection across five hospitals – a clinically relevant scenario where each client naturally maps to a different hospital.

4.2 Experimental Setup

Federated setup and training. We simulate a decentralised federated setting with $K = 4$ clients and $T = 10$ communication rounds where models are trained on the ID datasets (CIFAR10, HAM10000, APTOS, Camelyon17). For IID experiments (Table 2), training data is partitioned uniformly across clients. To exclude spurious effects such as label shift, we re-sample the OOD datasets to match the class proportions of their respective ID counterparts, thus isolating

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	**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**

the clinically relevant distribution shifts. For non-IID experiments on Camelyon17 (Table 3), each client corresponds to a different hospital and is evaluated on the held-out hospital. Clients exchange checkpoints at each round and retain all received checkpoints for agreement-based evaluation. We use FedAvg [21] as the aggregation strategy, SlimResNet18 for CIFAR-10, and ResNet18 [8] for HAM10000, APTOS, and Camelyon17. 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).

Evaluation metric. Following Baek et al. [1], we report Mean Absolute Error (MAE) between predicted and true OOD accuracy in percentage points (lower is better). We compute MAE at the last communication round, average across clients, and report mean and standard deviation over three random seeds.

Baselines. We compare the FedAgree strategies against Average Confidence (AC) [9] and Average Threshold Confidence (ATC) [5]. Both operate on individual models and require no additional checkpoints. We also include the original AotL baseline, which applies Eqs.(3) and (4) using only each client's own historical checkpoints. Code for reproducibility purposes is available at https://github.com/MLO-lab/FedAgree.

4.3 Experimental Results

Table 2 presents IID results across four benchmarks; FedAgree consistently outperforms both confidence-based baselines (AC, ATC) and AotL across all benchmarks. The advantage is particularly pronounced on medical imaging datasets where distribution shifts reflect 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% compared to AotL and 85% compared to AC. These results demonstrate that FedAgree is particularly well-suited for medical imaging applications where trained models

Table 3. Mean Absolute Error (MAE ↓) for Non-IID experiments on Camelyon. 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	**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**

must generalise across different acquisition devices, patient populations, and clinical sites.

The Non-IID Case. Table 3 reports non-IID results on Camelyon17-WILDS. In this realistic setting – where each client represents a different hospital – the benefits of cross-client checkpoints become even more pronounced, with all FedAgree strategies substantially outperforming AC and ATC. Heterogeneous client data yields checkpoints that capture complementary model behaviours, providing the diversity needed for more precise OOD estimation. FedAgree consistently outperforms all baselines, despite lower R^2 values (0.30–0.72) indicating weaker linear fits that suggest estimates should be treated with caution.

5 Discussion and Conclusion

In this paper, we proposed FedAgree, a method for agreement-based label-free OOD performance estimation in federated learning scenarios at minimal 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, respectively).

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. Furthermore, the coefficient of determination R^2 serves as a practical sanity check for the reliability of OOD estimates: higher R^2 values indicate a stronger linear fit and more precise estimates, while lower R^2 values suggest practitioners should interpret estimates with appropriate caution, despite FedAgree still outperforming the baselines. As a practical rule, if R^2 is low in safety-critical scenarios, we recommend falling back to manual performance estimation by labelling a representative set of images as a secure alternative.

Our findings suggest that the optimal strategy depends on the nature of the distribution shift. Selective strategies (Peer, Anchor Peer) work well under mild shifts where the linear assumption holds strongly, while Full – which maximises checkpoint diversity – performs best under severe domain shifts and is *always* better than local-only solutions.

Limitations and future work. 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) is interesting future work.

Acknowledgments. Work 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.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baek, C., Jiang, Y., Raghunathan, A., Kolter, J.Z.: Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. Advances in Neural Information Processing Systems **35**, 19274–19289 (2022)
2. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., 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)
3. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., et al.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science advances **8**(31), eabq6147 (2022)
4. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
5. Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H.: Leveraging unlabeled data to predict out-of-distribution performance. In: International Conference on Learning Representations (2022)
6. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: International Conference on Learning Representations (2021)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
10. Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N.C., Rotemberg, V., Halpern, A.C., Reiter, O., Carrera, C., Barreiro, A., Helba, B., et al.: Bcn20000: Dermoscopic lesions in the wild. Scientific data **11**(1), 641 (2024)
11. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: 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)
12. Karthik, Maggie, Dane, S.: Aptos 2019 blindness detection. https://kaggle.com/competitions/aptos2019-blindness-detection (2019), kaggle
13. Koebler, A., Decker, T., Thon, I., Tresp, V., Buettner, F.: Incremental uncertainty-aware performance monitoring with active labeling intervention. In: International Conference on Artificial Intelligence and Statistics. pp. 2188–2196. PMLR (2025)
14. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International conference on machine learning. pp. 5637–5664. PMLR (2021)
15. Lalitha, A., Shekhar, S., Javidi, T., Koushanfar, F.: Fully decentralized federated learning. In: Third workshop on bayesian deep learning (NeurIPS). vol. 12 (2018)
16. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5542–5550 (2017)
17. Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P.: Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:2108.13624 (2021)
18. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**(6) (2022)
19. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. Advances in neural information processing systems **31** (2018)
20. Lu, S., Nott, B., Olson, A., Todeschini, A., Vahabi, H., Carmon, Y., Schmidt, L.: Harder or different? a closer look at distribution shift in dataset reproduction. In: ICML Workshop on Uncertainty and Robustness in Deep Learning. vol. 5, p. 15 (2020)
21. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
22. Miller, J.P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P.W., Shankar, V., Liang, P., Carmon, Y., Schmidt, L.: 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)
23. Morteza, P., Li, Y.: Provable guarantees for understanding out-of-distribution detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 7831–7840 (2022)
24. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International conference on machine learning. pp. 10–18. PMLR (2013)
25. Peng, B., Luo, Y., Zhang, Y., Li, Y., Fang, Z.: Conjnorm: Tractable density estimation for out-of-distribution detection. In: The Twelfth International Conference on Learning Representations (2024)

26. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do cifar-10 classifiers generalize to cifar-10? arXiv preprint arXiv:1806.00451 (2018)
27. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future of digital health with federated learning. NPJ digital medicine **3**(1), 119 (2020)
28. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 180161 (2018)
29. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific data **10**(1), 41 (2023)
30. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4396–4415 (2022)