Task2Vec Readiness: Diagnostics for Federated Learning from Pre-Training Embeddings

Cristiano Mafuz[1], **Pedro Silva** [2], **Rodrigo Silva**[2]

[1]Programa de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil

[2]Departamento de Computação, Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil

cristiano.mafuz@aluno.ufop.edu.br

Abstract. *Federated Learning (FL) performance is highly sensitive to client heterogeneity, yet practitioners lack tools to anticipate training outcomes before training begins. We introduce Task2Vec Readiness, a pre-training diagnostic framework that embeds each client's local dataset into a numerical vector and derives scalar indices from the federation embedding space. Across four datasets, three federation sizes and two aggregation rules, Pearson and Spearman correlations between readiness indices and final FL performance frequently exceed* 0.9 *and remain significant after false-discovery-rate correction, establishing federation geometry as a proxy for collaborative training success that requires neither access to raw client data nor knowledge of the process that generated the partition.*

Resumo. *O desempenho da Aprendizagem Federada (FL) é altamente sensível à heterogeneidade dos clientes, mas os usuários carecem de ferramentas para antecipar os resultados do treinamento antes que ele comece. Introduzimos o Task2Vec Readiness, um framework de diagnóstico pré-treinamento que embute o conjunto de dados local de cada cliente em um vetor numérico e deriva índices escalares de prontidão a partir do espaço de embeddings da federação. Em quatro conjuntos de dados, três tamanhos de federação e dois algoritmos de agregação, as correlações de Pearson e Spearman entre os índices propostos e o desempenho final frequentemente superam* $0{,}9$ *e permanecem significativas após correção para múltiplas comparações, estabelecendo a geometria da federação como um proxy do sucesso do treinamento colaborativo que não exige acesso aos dados brutos dos clientes nem conhecimento do processo que gerou a partição.*

1. Introduction

Federated Learning (FL) [McMahan et al. 2017] has emerged as a central paradigm for collaborative model training under privacy and communication constraints [Zhang et al. 2021]. A necessary precondition for successful FL, however, is that the federation itself is well-formed: the participating clients must hold data distributions that are compatible enough to support collaborative optimization. This is not guaranteed. Client heterogeneity, data imbalance, and federation size can all undermine convergence and degrade final model quality [Yang et al. 2025]. Yet the decision of whether to federate, and

with whom, must typically be made before training begins and under privacy constraints that prevent direct inspection of client data. Without a principled way to assess federation viability in advance, practitioners are left with costly trial-and-error: committing to a federation, running a full training pipeline, observing failure, and iterating. Despite significant advances in aggregation and optimization [Xu et al. 2024, Yang et al. 2025], the field still lacks a pre-training diagnostic to answer the most fundamental question in federation formation: *given this set of clients, is collaborative training likely to succeed?*

In this context, we introduce *Task2Vec Readiness*, a framework that leverages Task2Vec embeddings [Achille et al. 2019] to derive a quantitative readiness index for federated learning. Our approach transforms each client's data distribution into a fixed-dimensional embedding via Fisher Information, and then evaluates unsupervised metrics of federation structure. In particular, we measure cohesion (average cosine similarity among client embeddings), dispersion (average distance from the federation centroid), and density (RBF-kernel similarity over pairwise Euclidean distances). These metrics can be computed before training, and together they form a readiness profile that anticipates how well the federation can support collaborative optimization. The key novelty lies in repurposing task embeddings from transfer learning into a diagnostic signal for distributed training under heterogeneity.

We conduct extensive experiments across four benchmark datasets (CIFAR-10 [Krizhevsky et al. 2009], FEMNIST [Albaseer et al. 2021], PathMNIST [Yang et al. 2023], BloodMNIST [Yang et al. 2023]), with client counts ranging from 10 to 30 and Dirichlet non-IID partitions spanning $\alpha \in \{0.05, \ldots, 5.0\}$. Correlation analyses consistently reveal significant Pearson and Spearman coefficients between readiness metrics and final model performance, frequently exceeding 0.9 across dataset$\times$client conditions. This validates readiness as a robust proxy for FL performance. Importantly, the results hold across different sources of heterogeneity, indicating that readiness captures structural properties of federations rather than dataset-specific artifacts.

Our contributions are threefold. First, we frame *federation readiness* as a pre-training diagnostic for the federation formation problem, assessing whether a given set of clients should collaborate at all, and distinguish it from client selection, which operates during training and presupposes that the federation already exists. Second, we repurpose Task2Vec embeddings, originally developed for task similarity in meta-learning, into this setting, and derive a family of geometric readiness indices (cohesion, dispersion, density, and the Cohesion–Dispersion Index) computable without access to raw client data.

2. Related Work

A substantial body of work characterizes client heterogeneity through statistics of the label distribution, including average label entropy [Lutz et al. 2025, Orlandi et al. 2023] and pairwise distribution distances such as Earth Mover's Distance and Jensen–Shannon divergence [Caldas et al. 2018, Solans et al. 2024]. These measures are cheap and interpretable, but they are blind to feature- and covariate-level structure: two clients with identical label marginals but disjoint feature distributions appear homogeneous under any label-only statistic, despite being difficult to federate. Our embedding-based indices operate in feature space and therefore capture heterogeneity that label statistics cannot.

Estimating transferability without full training has been studied through pairwise

metrics like LEEP and LogME. However, federation formation requires assessing mutual compatibility across K symmetric peers simultaneously, making Task2Vec's single symmetric embedding per client a more direct fit.

Much of the FL literature on "who participates" addresses client selection: choosing, at each round, a subset of available clients to sample, typically in cross-device settings [Pei et al. 2024]. This is a within-training decision that assumes a federation already exists. The complementary and earlier question is whether a set of parties should form a federation at all. This has received far less attention, and is the problem our readiness diagnostic targets.

3. Background

3.1. Federated Learning

The prototypical FL setting [McMahan et al. 2017] consists of a central server S and a set of distributed clients C, such that $|C| = K$, that jointly cooperate to solve a standard supervised learning task. Each client $c \in C$ has access to its own private training set $D_c = \{x_{c,i}, y_{c,i}\}_{i=1}^{n_c}$ where, $x_{c,i}$ denotes the i-th training sample of the c-th client, and $y_{c,i}$ denotes its corresponding label. The goal of FL is to train a global predictive model whose architecture and parameters $w \in \mathbb{R}_d$ are shared amongst all the clients and found to minimize $min_w \sum_{c=1}^{K} p_c L_c(w; D_c)$, where L_c is the local objective and $p_c \geq 0$ specifies the individual contribution of the client c such that $\sum_{c=1}^{K} p_c = 1$. Two possible configurations for p_c are $p_c = \frac{1}{K}$ or $p_c = \frac{n_c}{N}$, where $N = \sum_{c=1}^{K} n_c$.

The local objective function L_c usually is defined as the empirical risk calculated over the training set D_c sampled from the client's local data distribution $L_c(w; D_c) = \frac{1}{n_c} \sum_{i=1}^{n_c} l(w; (x_{c,i}, y_{c,i}))$, where l is an instance-level loss (e.g., cross-entropy loss or squared error in the case of classification or regression tasks, respectively). In its standard form, Federated Learning employs Federated Averaging (FedAvg) [McMahan et al. 2017], which operates over iterative rounds:

1. The server S selects a subset of clients $C^{(t)} \subseteq C$ and broadcasts the current global model $w^{(t)}$.
2. Each selected client updates $w^{(t)}$ locally on its data via stochastic gradient descent for multiple epochs, producing $\theta_c^{(t)}$.
3. The server aggregates updates using an aggregation function, typically a weighted average:

$$w^{(t+1)} = \sum_{c \in C^{(t)}} \frac{n_c}{N} w_c^{(t)}. \tag{1}$$

A known limitation of FedAvg under statistical heterogeneity is client drift: when local data distributions diverge significantly across clients, multiple steps of local SGD push each client's model towards its own local optimum, compromising the quality of the aggregated global model [Li et al. 2020]. To address this, [Li et al. 2020] proposed Federated Proximal (FedProx), which augments each client's local objective with a proximal term that penalises deviation from the current global model:

$$h_c(w; w^{(t)}) = L_c(w; D_c) + \frac{\mu}{2} \|w - w^{(t)}\|^2, \tag{2}$$

where $\mu \geq 0$ is a regularisation coefficient controlling the strength of the proximal penalty. Each client then minimises $h_c(w; w^{(t)})$ locally instead of L_c alone, keeping local updates closer to the global model and stabilising convergence under heterogeneous data distributions. When $\mu = 0$, FedProx reduces to FedAvg. The server aggregation step remains identical to Equation (1).

3.2. Task2Vec

In Task2Vec, a task is represented by a labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$. Then, the samples are passed through a pre-trained convolutional network, referred to as the probe network, and the representation of the task is captured by computing the diagonal of the Fisher Information Matrix (FIM) associated with its filter parameters.

Given a model with parameters w and output distribution $p_w(y \mid x)$, the FIM is the expected covariance of the scores (gradients of the log-likelihood) with respect to the model parameters. It can be defined as:

$$F = \mathbb{E}_{x,y\sim\hat{p}(x,y)}\Big[\nabla_w \log p_w(y \mid x)\, \nabla_w \log p_w(y \mid x)^\top\Big], \tag{3}$$

where $\hat{p}(x, y)$ denotes the empirical distribution of the dataset.

Since the full FIM may be prohibitively large depending on the probe network, [Achille et al. 2019] introduce two approximations: (i) only the diagonal entries are retained, that is, it assumes that correlations between different filters in the probe network are not important, and (ii) the Fisher information is averaged across all weights within the same filter, since these weights are usually not independent. The resulting representation of the task therefore has fixed dimensionality, equal to the number of filters in the probe network. The extraction process is summarized below (adapted from [Achille et al. 2019]):

Given a labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{K}$ and a fixed probe network P (feature extractor), we compute the embedding as follows:

1. **Freeze P, train the classifier head.** Cache features $z = P(x)$ and fit a linear classifier h_θ on top, leaving the feature extractor weights fixed.
2. **Estimate Fisher on the feature extractor.** Excluding the classifier, compute a *diagonal* Fisher proxy for the parameters w of P using the *Monte Carlo (score matching):* iterate minibatches, forward x, sample labels $y \sim p_w(y \mid x)$, compute the loss $\ell(x, y)$, backprop, and *accumulate* squared gradients per weight (an estimator of the diagonal Fisher).
3. **Filter-level aggregation (fixed dimensionality).** For each convolutional filter, average the diagonal Fisher values across the weights belonging to that filter, yielding a single scalar per filter.
4. **Embedding.** Concatenate the per-filter scalars (classifier excluded) to obtain the task embedding $\mathbf{v} = \phi(D; P) \in \mathbb{R}^m$, where m is the number of filters in P.

Broadly speaking, the FIM quantifies how sensitive the model's predictions are to small changes in its parameters. Directions in parameter space with high Fisher information correspond to features that are particularly important for solving the task, whereas directions with low values correspond to irrelevant or redundant features. By compressing this information into a fixed-length vector, Task2Vec creates an embedding that reflects

both the difficulty of the task (through the norm of the embedding) and its similarity to other tasks (through distances between embeddings).

4. Federation Readiness

Definition. Let $\mathcal{F} = \{D_1, D_2, \ldots, D_K\}$ denote a federation composed of K clients, each holding a local dataset D_c. Given a fixed probe network P, we compute the Task2Vec embedding of each client's dataset, denoted by

$$\mathbf{v}_c = \phi(D_c; P) \in \mathbb{R}^m, \quad c = 1, \ldots, K, \tag{4}$$

where $\phi(D_c; P)$ is the Task2Vec embedding vector defined in Section 3.2. The collection of client embeddings

$$V(\mathcal{F}; P) = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K\} \tag{5}$$

forms the *federation embedding space*, which provides a geometric representation of the heterogeneity and structure of the federation.

A *readiness index* is a scalar functional

$$R : V(\mathcal{F}; P) \mapsto \mathbb{R}, \tag{6}$$

that maps the set of client embeddings to a real value measuring the expected ability of the federation to support collaborative learning.

The rationale is that a federation with high readiness corresponds to embeddings that are compact, balanced, and diverse enough to provide global coverage of the underlying task distribution. Such federations are structurally aligned and therefore more likely to achieve efficient collaborative optimization under standard aggregation rules (e.g., FedAvg). Conversely, low readiness signals that the federation suffers from excessive heterogeneity, imbalance, or sparsity, making convergence difficult and final performance unreliable. In the next section, we present candidates for the Federation Readiness Index.

4.1. Readiness Indices

In this section, we present the tested readiness indices. First, we present the proposed *embedding-based indices*, derived directly from Task2Vec client embeddings. For comparison, we also present the federation entropy index, which captures statistical properties of the federation and has been used to characterize expected training difficulty [Caldas et al. 2018, Solans et al. 2024].

Cohesion. Measures the average cosine similarity between all pairs of client embeddings:

$$\text{Cohesion} = \frac{1}{K(K-1)} \sum_{i \neq j} \cos(\mathbf{v}_i, \mathbf{v}_j). \tag{7}$$

Intuitively, high cohesion means that clients share similar feature relevance profiles, suggesting they can collaborate effectively with minimal conflict.

Dispersion. Captures the average distance of client embeddings to the federation centroid, normalized by the embedding dimensionality:

$$\text{Dispersion} = \frac{1}{K} \sum_{i=1}^{K} \frac{\|\mathbf{v}_i - \bar{\mathbf{v}}\|_2}{\sqrt{|\mathbf{v}_i|}}. \tag{8}$$

High dispersion indicates that clients are widely spread in embedding space, a signal of strong heterogeneity that may slow convergence. Since the goal is to quantify how well the federation is structured, from here on we use the negative of the dispersion as a readiness index.

Density. Quantifies the overall compactness of the federation via an RBF kernel applied to embedding distances:

$$\text{Density} = \frac{1}{K(K-1)} \sum_{i \neq j} \exp\left(-\frac{\|\mathbf{v}_i - \mathbf{v}_j\|_2^2}{2\sigma^2} \right), \tag{9}$$

where σ is the median of pairwise distances. Federations with high density correspond to tightly packed embeddings, suggesting stable aggregation dynamics.

Cohesion–Dispersion Index (CDI). We define the Cohesion–Dispersion Index (CDI) as a weighted combination of cohesion and dispersion:

$$\text{CDI}_{\lambda,\gamma} = \lambda \cdot \text{Cohesion} + \gamma \cdot (-\text{Dispersion}), \tag{10}$$

where $\lambda, \gamma > 0$ are weighting coefficients.

This index balances the tendency of clients to align with one another (high cohesion) against the degree to which they are scattered in embedding space (low dispersion). A high CDI value therefore indicates a federation that is both well-aligned and structurally compact, properties expected to facilitate stable and efficient collaborative training. In this work, we set $\beta = 1$ and $\gamma = 1000$ to scale the dispersion term to the same order of magnitude as cohesion.

Average Entropy As a baseline for comparison, we adopt Average Entropy, a widely used measure of client heterogeneity in federated learning [Lutz et al. 2025, Orlandi et al. 2023]. It measures the average label entropy across clients:

$$\text{AvgEntropy} = \frac{1}{K} \sum_{i=1}^{K} H(p_i), \quad H(p_i) = -\sum_{c \in \mathcal{C}_i} p_{i,c} \log p_{i,c}. \tag{11}$$

where K is the number of clients, $\mathcal{C}_i$ is the set of classes present in client i's local dataset, $p_{i,c}$ is the proportion of samples belonging to class c in client i's local data, and $H(p_i)$ is the Shannon entropy of client i's label distribution. High entropy means that client data are locally balanced across classes, whereas low entropy reflects strong label imbalance.

5. Experimental Setup

We designed an experimental pipeline to evaluate the relationship between Task2Vec-based readiness and FL performance under varying degrees of data heterogeneity. All experiments sweep datasets, federation sizes, heterogeneity parameters, and aggregation strategies.

Datasets. We evaluate our approach on four benchmark datasets covering natural and medical image classification:

- **CIFAR-10** [Krizhevsky et al. 2009]: 60,000 RGB images (32×32) across 10 classes.

- **EMNIST-Letters**, used as a proxy for FEMNIST [Caldas et al. 2018]: grayscale handwritten character images. The natural writer-based partitioning of FEMNIST is *not* used; to keep the heterogeneity axis comparable across datasets, EMNIST-Letters is partitioned with the same label-wise Dirichlet procedure applied to the remaining datasets. Our results on this dataset therefore characterise synthetic label-shift heterogeneity on handwritten characters and do not speak to the natural writer-induced heterogeneity of the original FEMNIST benchmark.
- **PathMNIST** and **BloodMNIST** [Yang et al. 2023]: Medical imaging datasets from MedMNIST v2. PathMNIST contains 100,000 histopathology images of colorectal tissue (9 classes); BloodMNIST contains 17,092 microscopic blood-cell images (8 classes).

For CIFAR-10 and MedMNIST datasets, synthetic non-IID splits are generated by partitioning data across clients via a Dirichlet distribution over class labels, controlled by α. Low α yields highly non-IID distributions; high α approaches IID. All datasets, including EMNIST-Letters, are partitioned with the same Dirichlet procedure over class labels.

Preprocessing. All images are normalised by dataset-specific mean and standard deviation. For Task2Vec computation, all samples are converted to 3-channel RGB and resized to 224×224 to match the probe network's requirements.

Baseline Model. All federations train a ResNet34 [He et al. 2016, PyTorch 2025] initialised from scratch; this is distinct from the probe network used for Task2Vec, which is ImageNet-pretrained. Single-channel inputs are replicated to three channels in the preprocessing transform rather than by adapting the first convolutional layer, so the architecture is identical across datasets except for the final fully-connected layer, which is sized per dataset. All other architecture choices are fixed, ensuring that performance variation is attributable to data distribution and federation structure rather than model capacity.

Federation Design. Federations vary along two axes: the number of clients K and the degree of statistical heterogeneity α.

- **Number of clients**: $K \in \{10, 20, 30\}$.
- **Optimiser**: SGD with momentum 0.9.
- **Learning rate**: 0.01.
- **Communication rounds**: $T = 20$.
- **Aggregation strategies**: FedAvg [McMahan et al. 2017] and FedProx [Li et al. 2020], both with full client participation (`fraction_fit`$= 1.0$) and random half-client evaluation (`fraction_evaluate`$= 0.5$). For FedProx, the proximal regularisation coefficient is set to $\mu = 0.01$.
- **Local training**: $E = 1$ local epoch per round, batch size $B = 32$.
- **Heterogeneity**: $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.5, 1.0, 2.0, 5.0\}$ for CIFAR-10 and MedMNIST datasets.

Evaluation Metrics. Model performance is assessed using Top-1 accuracy for CIFAR-10 and FEMNIST, and macro-averaged AUC for PathMNIST and BloodMNIST to account for class imbalance. Final results are reported on a centralised test set held out exclusively for evaluation after training.

Task2Vec Readiness Computation. Before starting federated training, Task2Vec embeddings are computed for each client's dataset using ResNet34 pre-trained on ImageNet (`IMAGENET1K_V1`) as probe network P. Embedding parameters: maximum 1,000

samples per client, 6 skipped early layers, embedding dimension fixed at 1,000 filters. For single-channel inputs, images are replicated to 3 channels before embedding.

Distributional Baselines. To contextualise Task2Vec readiness, as mentioned before, we compute the Average Entropy (AvgEntropy) baseline that operates exclusively on client label distributions.

Reproducibility. Experiments use PyTorch and the Flower framework for FL orchestration, with the AWS reference implementation of Task2Vec.[1] Exact library versions, hyperparameters, hardware specification, and the scripts that regenerate every table and figure in this paper are available at `https://github.com/CristianoMafuz/task2vec_federated_learning`.

5.1. Statistical Protocol

Each experimental cell is identified by the tuple (dataset, aggregation algorithm, number of clients K, Dirichlet parameter α). For FedAvg, two independent federations—differing in the partition seed and therefore in client composition—were generated per cell; for FedProx, a single federation per cell was generated. The asymmetry is a consequence of the experimental schedule and is reported explicitly rather than absorbed into the tables.

Correlations are computed over the $|\{\alpha\}| = 8$ heterogeneity levels, with readiness and performance first averaged across the independent federations available at each α. The unit of analysis is therefore the heterogeneity level, and $n = 8$ for every cell in Tables 2 and 3, except BloodMNIST/FedAvg/$K = 20$, where one heterogeneity level is missing and $n = 7$. Sample sizes are reported explicitly for every test. Because twelve tests are reported per index and per algorithm, all p-values are corrected for multiple comparisons using the Benjamini–Hochberg procedure at a false discovery rate of 0.05; the values reported in parentheses are the adjusted ones. All tests are two-sided.

6. Results and Discussion

Table 1 reports raw final performance values (Top-1 accuracy for CIFAR-10/FEMNIST, macro-AUC for PathMNIST/BloodMNIST) across all combinations of dataset, algorithm, K, and α. Two patterns emerge. First, FedProx outperforms FedAvg in the majority of configurations, with the effect most pronounced for CIFAR-10. Second, this advantage is not universal: for FEMNIST and PathMNIST, FedAvg occasionally outperforms FedProx at the most extreme heterogeneity level ($\alpha = 0.05$), and for BloodMNIST, FedAvg becomes competitive again near the IID regime ($\alpha \geq 2.0$). This suggests that the benefit of FedProx's proximal regularisation is most valuable in moderate-to-high heterogeneity regimes, and its overhead is not always justified at the extremes.

Tables 2 and 3 report Pearson and Spearman correlations between Task2Vec-based readiness indices and final FL performance, computed over the eight Dirichlet heterogeneity levels for each dataset$\times K$ configuration under both FedAvg and FedProx. Statistical significance is assessed at $\alpha = 0.05$ (bolded cells; two-sided test).

Cohesion is significant in every configuration after correction, and negative dispersion in nine of twelve under FedAvg, with the exceptions concentrated in PathMNIST. The indices retain their association under FedProx, indicating that the signal is not specific

[1] `https://github.com/awslabs/aws-cv-task2vec`. Accessed: 2025-09-23.

Table 1. Final performances by K, algorithm, and Dirichlet α.

Algorithm	K	Dirichlet α 0.05	0.1	0.2	0.3	0.5	1.0	2.0	5.0
CIFAR-10 — final performance (Top-1 Accuracy)									
FedAvg	10	0.445	0.494	0.561	0.603	0.625	0.646	0.646	0.668
FedProx	10	**0.538**	**0.561**	**0.650**	**0.688**	**0.716**	**0.730**	**0.756**	**0.757**
FedAvg	20	0.428	0.481	0.517	0.530	0.571	0.588	0.599	0.606
FedProx	20	**0.526**	**0.545**	**0.636**	**0.596**	**0.661**	**0.689**	**0.699**	**0.708**
FedAvg	30	0.404	0.458	0.498	0.506	0.538	0.557	0.566	0.565
FedProx	30	**0.486**	**0.477**	**0.560**	**0.570**	**0.637**	**0.630**	**0.652**	**0.671**
FEMNIST — final performance (Top-1 Accuracy)									
FedAvg	10	0.838	0.871	0.894	0.902	0.904	0.905	0.906	0.906
FedProx	10	**0.849**	**0.887**	**0.915**	**0.931**	**0.941**	**0.943**	**0.946**	**0.943**
FedAvg	20	**0.827**	0.880	0.891	0.895	0.899	0.902	0.904	0.903
FedProx	20	0.819	**0.887**	**0.932**	**0.931**	**0.938**	**0.941**	**0.941**	**0.943**
FedAvg	30	**0.827**	0.872	0.888	0.894	0.896	0.898	0.900	0.918
FedProx	30	0.820	**0.905**	**0.923**	**0.932**	**0.936**	**0.939**	**0.941**	**0.942**
PathMNIST — final performance (Macro-AUC)									
FedAvg	10	**0.906**	0.943	0.953	0.952	0.959	0.960	0.964	0.964
FedProx	10	0.878	**0.948**	**0.973**	**0.973**	**0.967**	**0.968**	**0.983**	**0.977**
FedAvg	20	0.915	**0.952**	0.952	0.954	0.960	0.962	0.957	0.962
FedProx	20	**0.948**	0.947	**0.963**	**0.975**	**0.968**	**0.981**	**0.968**	**0.969**
FedAvg	30	**0.910**	0.943	0.950	0.949	0.954	0.956	0.956	0.955
FedProx	30	0.900	**0.960**	**0.963**	**0.970**	**0.969**	**0.974**	**0.968**	**0.967**
BloodMNIST — final performance (Macro-AUC)									
FedAvg	10	0.923	**0.959**	0.979	0.982	0.987	0.989	0.990	0.991
FedProx	10	**0.926**	0.934	**0.985**	**0.992**	**0.995**	**0.996**	**0.995**	**0.996**
FedAvg	20	0.915	**0.961**	0.967	0.971	0.978	0.980	0.982	0.987
FedProx	20	**0.926**	0.938	**0.988**	**0.990**	**0.992**	**0.988**	**0.993**	**0.995**
FedAvg	30	0.919	0.937	0.962	0.968	0.973	0.978	**0.979**	**0.982**
FedProx	30	**0.953**	**0.950**	**0.980**	**0.977**	**0.987**	**0.983**	0.954	0.979

Table 2. Pearson correlation between readiness and final FL performance

Dataset	Alg. (K)	Cohesion	-Dispersion	Density	CDI	Avg Entropy
BloodMNIST	FedAvg ($K=10$)	**0.997(0.000)**	**0.886(0.006)**	-0.638(0.431)	**0.955(0.001)**	**0.873(0.008)**
	FedAvg ($K=20$)	**0.986(0.000)**	**0.879(0.012)**	-0.484(0.556)	**0.947(0.002)**	**0.871(0.014)**
	FedAvg ($K=30$)	**0.950(0.000)**	**0.881(0.006)**	-0.657(0.431)	**0.948(0.001)**	**0.936(0.002)**
	FedProx ($K=10$)	**0.893(0.003)**	**0.721(0.044)**	-0.484(0.224)	**0.893(0.003)**	**0.816(0.013)**
	FedProx ($K=20$)	**0.970(0.000)**	**0.742(0.035)**	-0.011(0.979)	**0.970(0.000)**	**0.774(0.024)**
	FedProx ($K=30$)	0.623(0.099)	0.278(0.505)	0.323(0.435)	0.623(0.099)	0.411(0.312)
CIFAR-10	FedAvg ($K=10$)	**0.994(0.000)**	**0.962(0.001)**	-0.137(0.922)	**0.987(0.000)**	**0.974(0.000)**
	FedAvg ($K=20$)	**0.975(0.000)**	**0.982(0.000)**	-0.076(0.922)	**0.997(0.000)**	**0.987(0.000)**
	FedAvg ($K=30$)	**0.976(0.000)**	**0.978(0.000)**	-0.438(0.556)	**0.996(0.000)**	**0.977(0.000)**
	FedProx ($K=10$)	**0.990(0.000)**	**0.905(0.002)**	0.201(0.634)	**0.990(0.000)**	**0.960(0.000)**
	FedProx ($K=20$)	**0.900(0.002)**	**0.966(0.000)**	-0.013(0.975)	**0.900(0.002)**	**0.958(0.000)**
	FedProx ($K=30$)	**0.877(0.004)**	**0.961(0.000)**	-0.580(0.132)	**0.878(0.004)**	**0.963(0.000)**
FEMNIST	FedAvg ($K=10$)	**0.968(0.000)**	**0.881(0.006)**	-0.042(0.922)	**0.909(0.002)**	**0.872(0.008)**
	FedAvg ($K=20$)	**0.961(0.000)**	**0.934(0.002)**	-0.119(0.922)	**0.943(0.001)**	**0.858(0.010)**
	FedAvg ($K=30$)	**0.952(0.000)**	**0.940(0.002)**	0.074(0.922)	**0.944(0.001)**	**0.914(0.004)**
	FedProx ($K=10$)	**0.976(0.000)**	**0.965(0.000)**	0.112(0.791)	**0.976(0.000)**	**0.911(0.002)**
	FedProx ($K=20$)	**0.986(0.000)**	**0.996(0.000)**	-0.568(0.142)	**0.986(0.000)**	**0.868(0.005)**
	FedProx ($K=30$)	**0.921(0.001)**	**0.932(0.001)**	-0.202(0.631)	**0.921(0.001)**	**0.847(0.008)**
PathMNIST	FedAvg ($K=10$)	**0.925(0.001)**	0.677(0.065)	-0.082(0.922)	**0.790(0.020)**	**0.818(0.014)**
	FedAvg ($K=20$)	**0.940(0.001)**	0.682(0.065)	-0.566(0.431)	**0.818(0.014)**	**0.793(0.019)**
	FedAvg ($K=30$)	**0.982(0.000)**	0.700(0.064)	-0.584(0.431)	**0.868(0.006)**	**0.824(0.014)**
	FedProx ($K=10$)	**0.938(0.001)**	0.573(0.138)	0.125(0.769)	**0.938(0.001)**	0.670(0.069)
	FedProx ($K=20$)	**0.781(0.022)**	0.704(0.051)	-0.571(0.139)	**0.781(0.022)**	**0.716(0.046)**
	FedProx ($K=30$)	**0.928(0.001)**	0.603(0.113)	-0.619(0.102)	**0.927(0.001)**	0.697(0.055)

*bold indicates $p < 0.05$ after Benjamini–Hochberg correction; correlations computed over 8 Dirichlet levels (7 for BloodMNIST FedAvg $K = 20$, one level absent).

to a single aggregation rule. We note that the Pearson coefficients are high, but we do not read this as evidence of an intrinsically linear readiness–performance law: both quantities are monotone functions of the same underlying heterogeneity, and monotone covariates sampled on a regular grid produce high linear correlations as a matter of course.

Density is significant in none of the twelve FedAvg configurations, and its sign is negative in eleven of them—the opposite of what the compactness rationale predicts. We report it as a failed index rather than an inconclusive one. Its per-federation median bandwidth also makes values incomparable across federations, which alone would disqualify it as a cross-federation diagnostic.

Negative Dispersion is a strong predictor in most settings, especially for larger K where the centroid estimate is more stable. Cohesion and CDI yield virtually identical correlations across all configurations, a consequence of the CDI weighting: despite the = 1000 scaling intended to bring dispersion to the same order of magnitude as cohesion, the dispersion term remains dominated in practice, reducing CDI to a near-linear transformation of cohesion alone.

Table 3. Spearman correlation between readiness and final FL performance

Dataset	Alg. (K)	Cohesion	-Dispersion	Density	CDI	Avg Entropy
BloodMNIST	FedAvg ($K=10$)	**1.000(0.000)**	**1.000(0.000)**	-0.452(0.496)	**1.000(0.000)**	**1.000(0.000)**
	FedAvg ($K=20$)	**1.000(0.000)**	**0.964(0.001)**	-0.571(0.496)	**0.964(0.001)**	**1.000(0.000)**
	FedAvg ($K=30$)	**1.000(0.000)**	**1.000(0.000)**	-0.643(0.342)	**1.000(0.000)**	**1.000(0.000)**
	FedProx ($K=10$)	**0.929(0.001)**	**0.905(0.002)**	-0.214(0.610)	**0.929(0.001)**	**0.929(0.001)**
	FedProx ($K=20$)	**0.929(0.001)**	**0.929(0.001)**	0.071(0.867)	**0.929(0.001)**	**0.929(0.001)**
	FedProx ($K=30$)	0.452(0.260)	0.452(0.260)	0.333(0.420)	0.452(0.260)	0.452(0.260)
CIFAR-10	FedAvg ($K=10$)	**1.000(0.000)**	**1.000(0.000)**	-0.286(0.591)	**1.000(0.000)**	**1.000(0.000)**
	FedAvg ($K=20$)	**1.000(0.000)**	**1.000(0.000)**	-0.143(0.736)	**1.000(0.000)**	**1.000(0.000)**
	FedAvg ($K=30$)	**0.976(0.000)**	**0.976(0.000)**	-0.333(0.560)	**0.976(0.000)**	**0.976(0.000)**
	FedProx ($K=10$)	**1.000(0.000)**	**0.976(0.000)**	-0.143(0.736)	**1.000(0.000)**	**1.000(0.000)**
	FedProx ($K=20$)	**0.976(0.000)**	**1.000(0.000)**	0.333(0.420)	**0.976(0.000)**	**0.976(0.000)**
	FedProx ($K=30$)	**0.952(0.000)**	**0.952(0.000)**	-0.476(0.233)	**0.952(0.000)**	**0.952(0.000)**
FEMNIST	FedAvg ($K=10$)	**1.000(0.000)**	**0.762(0.028)**	-0.190(0.711)	**0.976(0.000)**	**1.000(0.000)**
	FedAvg ($K=20$)	**0.976(0.000)**	**0.976(0.000)**	-0.429(0.496)	**0.976(0.000)**	**0.976(0.000)**
	FedAvg ($K=30$)	**1.000(0.000)**	**1.000(0.000)**	0.381(0.528)	**1.000(0.000)**	**1.000(0.000)**
	FedProx ($K=10$)	**0.929(0.001)**	**0.976(0.000)**	0.071(0.867)	**0.929(0.001)**	**0.929(0.001)**
	FedProx ($K=20$)	**0.976(0.000)**	**0.905(0.002)**	-0.310(0.456)	**0.976(0.000)**	**0.976(0.000)**
	FedProx ($K=30$)	**0.976(0.000)**	**0.976(0.000)**	0.119(0.779)	**0.976(0.000)**	**1.000(0.000)**
PathMNIST	FedAvg ($K=10$)	**0.976(0.000)**	**0.952(0.000)**	-0.500(0.496)	**0.952(0.000)**	**0.976(0.000)**
	FedAvg ($K=20$)	**0.905(0.002)**	**0.905(0.002)**	-0.667(0.342)	**0.905(0.002)**	**0.905(0.002)**
	FedAvg ($K=30$)	**0.881(0.004)**	**0.881(0.004)**	-0.667(0.342)	**0.881(0.004)**	**0.881(0.004)**
	FedProx ($K=10$)	**0.762(0.028)**	**0.738(0.037)**	0.143(0.736)	**0.762(0.028)**	**0.762(0.028)**
	FedProx ($K=20$)	0.667(0.071)	0.571(0.139)	-0.286(0.493)	0.667(0.071)	0.667(0.071)
	FedProx ($K=30$)	0.595(0.120)	0.500(0.207)	-0.500(0.207)	0.595(0.120)	0.595(0.120)

*bold indicates $p < 0.05$ after Benjamini–Hochberg correction; correlations computed over 8 Dirichlet levels (7 for BloodMNIST FedAvg $K=20$, one level absent).

7. Conclusion

We posed the federation formation problem as: Given a set of clients, can the viability of collaborative training be assessed before it begins? We have answered it with

Task2Vec Readiness, a pre-training diagnostic that embeds each client's data into a Fisher-information vector and summarizes the federation's geometry through scalar indices. Our central finding is that this geometry is predictive: cohesion and negative dispersion track final performance closely and consistently, and they do so under both FedAvg and FedProx, indicating that the signal reflects properties of the federation rather than of any particular aggregation rule.

The comparison with Average Entropy clarifies what the embedding-based view does and does not add. Under the label-shift heterogeneity that this study induces, the two approaches largely agree, and we find no configuration in which the Fisher-information geometry is the only usable signal. The case for embeddings rests on a regime this design never instantiates—heterogeneity in feature space under fixed label marginals—and testing that regime is the natural next step rather than a claim we can make here.

This study has several limitations. The probe network is fixed (ResNet34 pretrained on ImageNet), leaving open how probe choice affects diagnostic quality across domains. The experiments are restricted to image classification and small-to-moderate federation scales, so the behavior of readiness indices in cross-silo NLP or time-series federations is untested. The CDI relies on a fixed weighting ($\beta = 1$, $\gamma = 1000$) chosen to balance the magnitude of its two terms rather than tuned, and its sensitivity to this choice is not characterized. Density, in turn, remains inconsistent across configurations, capturing local cluster compactness rather than global federation geometry. Each of these points lead to a direction for future work: alternative probe architectures, applications beyond image classification, larger federation scales, principled weighting for the CDI, and refined density definitions that better capture both local clustering and global alignment.

Finally, a more fundamental step remains. Our results establish a correlation between readiness and performance, but not yet a decision rule. The natural next step is to turn readiness into an actionable criterion, a threshold for whether to federate, or an objective for choosing among candidate federations, and to determine how reliably such a rule transfers across the settings above. Bridging this gap, from diagnostic signal to formation decision, is what would make readiness genuinely useful in practice.

Acknowledgements

The authors thank CNPq, FAPEMIG, and PROPPI/UFOP for the financial support provided to this research project.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. (2019). Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439.

Albaseer, A., Abdallah, M., Al-Fuqaha, A., and Erbad, A. (2021). Client selection approach in support of clustered federated learning over wireless edge networks.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. (2018). Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Lutz, A., Steidl, G., Müller, K., and Samek, W. (2025). Optimizing federated learning by entropy-based client selection. In *2025 3rd International Conference on Federated Learning Technologies and Applications (FLTA)*, pages 218–227.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Orlandi, F. C., Dos Anjos, J. C. S., Leithardt, V. R. Q., De Paz Santana, J. F., and Geyer, C. F. R. (2023). Entropy to mitigate non-iid data problem on federated learning for the edge intelligence environment. *IEEE Access*, 11:78845–78857.

Pei, J., Liu, W., Li, J., Wang, L., and Liu, C. (2024). A review of federated learning methods in heterogeneous scenarios. *IEEE Transactions on Consumer Electronics*, 70(3):5983–5999.

PyTorch (2025). torchvision.models.resnet34. `https://pytorch.org/vision/main/models/generated/torchvision.models.resnet34.html`. Accessed: 2025-09-23.

Solans, D., Heikkila, M., Vitaletti, A., Kourtellis, N., Anagnostopoulos, A., Chatzigiannakis, I., et al. (2024). Non-iid data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. *arXiv preprint arXiv:2411.12377*.

Xu, S., Xia, H., Zhang, R., Liu, P., and Fu, Y. (2024). Fednor: A robust training framework for federated learning based on normal aggregation. *Information Sciences*, 684:121274.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. (2023). Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

Yang, Z., Zhang, Y., Li, C., Cheung, Y.-m., Han, B., and Yuan, Y. (2025). Fedgps: Statistical rectification against data heterogeneity in federated learning. In Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., and Chen, N., editors, *Advances in Neural Information Processing Systems*, volume 38, pages 94731–94772. Curran Associates, Inc.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., and Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, 216:106775.