FEDGME: FEDERATED GAUSSIAN MIXTURE-BASED LEARNING FOR CLINICAL EEG CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) enables collaborative model training across hospitals without sharing sensitive local patient data. In clinical EEG, however, label distributions differ strongly across sites, and rare classes, such as uncommon seizure types, may be missing from a hospital altogether. Weight-sharing methods such as FedAvg are costly, scale with model size and are biased toward majority classes under label skew, while prototype-sharing methods such as FedPCL summarise each class by a single point, which collapses the within-class variability intrinsic to EEG and carries little about classes a site has never seen. We propose FedGME (Federated Gaussian Mixture-based Learning for EEG), in which clients exchange low-rank Gaussian components per class instead of weights or prototypes. The server keeps the components of all sites side by side, so each class is shared as a mixture that preserves its spread and can be sampled, allowing clients to learn classes absent from their own data. Built on a pretrained EEG model and evaluated on three corpora from the TUH EEG Corpus, FedGME lets clients recognise classes they never observed, raising their balanced accuracy to 0.15–0.30, where prototype methods stay near zero and FedAvg reaches at most 0.10. On the two multiclass tasks, this improves overall balanced accuracy by 15.5% on average relative to FedAvg (up to $+0.060$ absolute), while transmitting only 0.2%–1.3% of FedAvg's payload and no model weights. Replacing FedPCL's prototype-wise contrastive loss with an established spread-preserving contrastive loss, weighted by patient identity, prevents the within-class collapse of prototype training and raises effective rank by up to $4.2\times$ over FedPCL at the same payload. On the balanced binary task, where no class is absent, FedGME performs on par with FedAvg. Overall, sharing class distributions instead of model weights offers a communication-efficient strategy for mitigating label skew in federated biomedical signal analysis.

1 INTRODUCTION

Federated learning (FL) is a distributed learning paradigm in which clients, such as hospitals, learn collaboratively without sharing their data (McMahan et al., 2017). Despite its growing adoption in Computer Vision (Liu et al., 2020), Natural Language Processing (Chen et al., 2019), and autonomous driving (Nguyen et al., 2022), FL remains underexplored for multiclass classification of clinical EEG, where heterogeneous site distributions and limited annotations pose domain-specific difficulties.

In clinical EEG, multiclass classification matters: treatment decisions may depend in part on seizure type (Glauser et al., 2013). It is, however, particularly challenging in federated settings. EEG has a low signal-to-noise ratio (Roy et al., 2019), encodes substantial subject-specific information (Wang et al., 2020; Özdenizci et al., 2020), and is susceptible to artifacts such as eye movements (Jiang et al., 2019), while classes are severely imbalanced and their distributions differ across hospitals (non-IID). Under such conditions, Federated Averaging (FedAvg) generalises poorly (McMahan et al., 2017; Zhao et al., 2018), and its repeated exchange of full model parameters becomes a communication bottleneck as models grow (Qin et al., 2024).

To address these challenges, federated contrastive learning (FCL) aligns class-wise representations across clients (Li et al., 2021; Seo et al., 2024), while other approaches reduce communication by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

exchanging compact signals such as soft predictions (Jeong et al., 2018; He et al., 2020) or class prototypes (Tan et al., 2022a); FedPCL (Tan et al., 2022b) combines both by applying contrastive objectives to class prototypes. For EEG, however, both can discard important within-class structure. Supervised contrastive learning can drive the samples of a class toward a single point, suppressing the subject-, device- and context-related variability intrinsic to biomedical signals (Khosla et al., 2020; Graf et al., 2021); similar concerns about excessive alignment arise in general FL (Seo et al., 2024). Prototypes impose the same constraint by representing each class as one point (Snell et al., 2017; Allen et al., 2019), or, in multi-prototype variants, as a few discrete points (Qiao et al., 2024).

Moreover, FCL methods typically evaluate clients only on local labels, although rare classes, such as uncommon seizure types, may appear at only a few hospitals; lacking positive examples, standard classifiers suppress them despite needing to recognise them for new patients. Pu et al. (2024) handle such unseen categories with Gaussian components but rely on full-parameter exchange, whereas FCL methods that avoid model exchange freeze the pretrained encoder (Tan et al., 2022b). Unfreezing it lets each client fine-tune on its own data, which distorts pretrained features (Kumar et al., 2022) and can leave client representation spaces inconsistent (Zhou et al., 2025), a mismatch that prototype updates alone cannot correct.

We address locally absent classes, single-point prototypes, representation collapse and encoder drift with a framework in which clients exchange class-level distributions rather than model parameters. Each client summarises every class as a low-rank Gaussian component; the server keeps these components side by side, so each class is shared as a mixture over sites rather than a single averaged point, and the mixture can be sampled to learn classes absent from a client's data. Institutions thus share both the mean and the spread of each class, not only its average, without sharing data or weights. Our contributions are:

- **Recognising classes a client has never observed, without exchanging weights.** Clients sample synthetic features for their absent classes from the shared Gaussian mixture (synthetic replay), while the encoder directions that carry these classes stay tied to the frozen pretrained reference (absent-subspace anchoring). Balanced accuracy on such classes rises to 0.15–0.30 (at most 0.10 for FedAvg, near zero for prototype methods), and overall balanced accuracy on the multiclass tasks improves on average by 15.5% relative to FedAvg.
- **Class distributions, not points, as the shared state.** The server keeps one low-rank Gaussian component per client and class instead of an averaged prototype, so each class retains its spread across sites, at 0.2%–1.3% of FedAvg's traffic and independent of model size.
- **Identifying and removing within-class collapse in prototype-based FL.** Rather than proposing a new objective, we reformulate FedPCL's loss with established supervised contrastive and class-conditional spreading losses (Khosla et al., 2020; Chen et al., 2022), weighted by patient identity; at the same payload, this raises effective rank by up to $4.2\times$.
- **A shared frozen reference for all client encoders.** Pretrained-feature and absent-subspace anchoring tie every client's encoder to the same pretrained model, so that components estimated at one site remain meaningful at another.

2 Related Work

We review three areas closely related to our work: federated learning for clinical EEG; contrastive representation learning, which provides the basis for training the client encoders; and prototype- and distribution-based federated learning, which, similar to our approach, communicates class-level summaries rather than model parameters.

2.1 Federated learning for clinical EEG

Early FL work on EEG federated small networks operating on spatial covariance matrices for motor-imagery BCI (Ju et al., 2020), whereas recent systems train local networks and aggregate their parameters (Ding et al., 2023; Lim et al., 2025; Arulmurugan et al., 2026), handling heterogeneous patients and sites through adaptive aggregation (Chan et al., 2024), personalization (Saemaldahr & Ilyas, 2023; Baghersalimi et al., 2022; Lou et al., 2025), decentralized optimization (Ding et al., 2023; Baghersalimi et al., 2024) and privacy mechanisms (Arulmurugan et al., 2026). Most of this

work, however, is limited to binary classification (Saemaldahr & Ilyas, 2023; Baghersalimi et al., 2022; Lou et al., 2025), although treatment choices require distinguishing several conditions, such as seizure types (Glauser et al., 2013), and FCL or prototype methods such as FedPCL have not been evaluated on such tasks.

2.2 Supervised and federated contrastive representation learning

Supervised contrastive learning attracts same-class and separates different-class samples (Khosla et al., 2020); under standard assumptions its optimum collapses each class onto a simplex vertex (Graf et al., 2021), discarding within-class geometry. Spread-preserving objectives add class-conditional repulsion to retain it (Chen et al., 2022). This matters for EEG, where much of the within-class variation comes from patients and recording conditions and representations strongly encode patient identity (Özdenizci et al., 2020): a useful representation should keep this spread while organising classes by structure shared across patients. FCL applies such objectives to mitigate client drift under non-IID data: the MOON approach contrasts local representations with global and previous local models (Li et al., 2021), and later methods address local collapse, label skew or feature heterogeneity (Seo et al., 2024; Huang et al., 2025). These methods rely on global model aggregation; we apply contrastive learning without exchanging parameters, so the communicated statistics themselves must remain comparable across independently adapted clients.

2.3 Prototype and distribution-based federated learning

Prototype-based FL exchanges class representatives instead of parameters: FedProto aggregates local prototypes into global ones (Tan et al., 2022a), FedPCL applies prototype-wise contrastive learning on fixed pretrained backbones (Tan et al., 2022b), and MP-FedCL represents a class by several k-means centroids (Qiao et al., 2024); directly exchanged prototypes, however, suffer from inconsistent client representation spaces (Zhou et al., 2025). For classes absent from a client, FedLC calibrates logits (Zhang et al., 2022), FedVLS distils vacant classes from a global model (Guo et al., 2025), CCVR samples virtual features from class-conditional Gaussians to recalibrate the classifier after training (Luo et al., 2021), and Pu et al. (2024) use Gaussian components but exchange full model parameters. Like CCVR, we sample Gaussian features; however, we draw them from client-specific components rather than aggregated class statistics, during training rather than after it, and only for classes a client lacks. An additional distinction from both CCVR and Pu et al. (2024) is that no global model is exchanged: each client's encoder adapts locally, so we also preserve the encoder subspace of absent classes to keep components estimated at other sites meaningful.

In summary, federated EEG methods rely on parameter exchange, prototype methods reduce each class to a point, and methods for locally absent classes depend on a shared global model. Our approach, FedGME, instead exchanges class-level distributions that preserve within-class spread and can be sampled to learn classes a site has never observed.

3 Methodology

Figure 1 outlines one communication round of FedGME. After fixing the setting (Sec. 3.1), the subsections follow its panels: the class summaries clients send (a, Sec. 3.2), the shared bank the server assembles (b, Sec. 3.3), the objective clients minimise against it (c–e, Sec. 3.4) and the treatment of locally absent classes (f, Sec. 3.5); Sec. 3.6 gives the training schedule. Each subsection states the concept and its rationale; definitions, estimation procedures and edge cases are in the matching parts of Appendix A.

3.1 Setting and overview

We consider K clients holding disjoint sets of patients recorded under the same protocol. Client k observes the classes $\mathcal{C}_k \subseteq \mathcal{C}$; the classes in $\mathcal{C} \setminus \mathcal{C}_k$ are *locally absent*, meaning the client holds no examples of them at all. For a window x the encoder gives $h = f_k(x) \in \mathbb{R}^{d_h}$, the projector the unit vector $z = p_k(h)/\|p_k(h)\|_2 \in \mathbb{S}^{d_z-1}$, and the linear probe $\hat{y} = g_k(h)$, where f_k is a BIOT encoder (Yang et al., 2023) with $d_h = 256$ and $d_z = 64$, initialised from a public checkpoint pretrained entirely outside the TUH corpus to rule out data leakage (Appendix B.2), p_k an MLP projector and

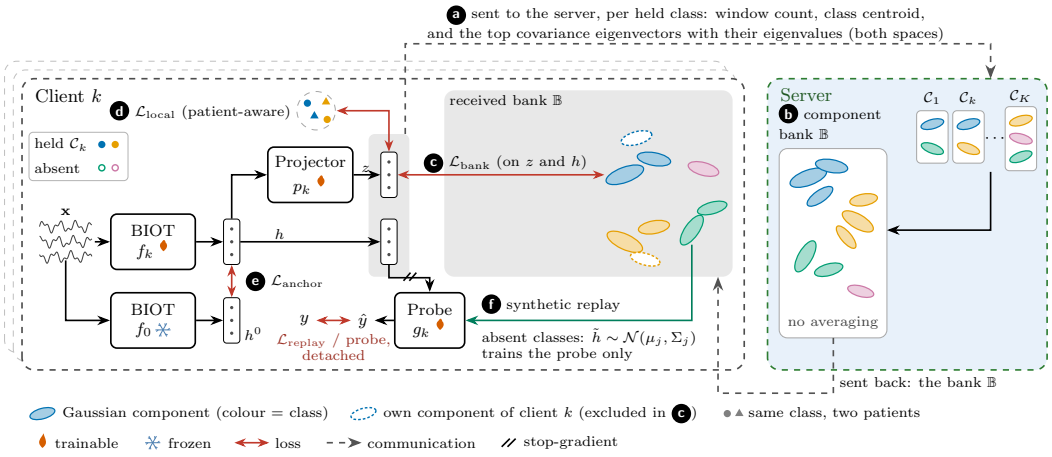


Figure 1: Overview of one communication round of FedGME. (a) Each client summarises every class it holds as a Gaussian component and sends it to the server. (b) The server collects the components into a shared bank without averaging them and returns it. (c)–(e) Each client trains against the bank: its representations are pulled toward the other clients' components of the same class (c), a patient-aware contrastive term organises the local batch (d), and representations stay anchored to the frozen pretrained encoder (e). (f) Classes the client never observes are learned from components sampled out of the bank. Separate banks are kept for z and h; one is shown.

g_k a linear probe. We refer to h and z as the representations of a window and write $\hat{h} = h/\|h\|_2$, since every statistic and score below is computed on ℓ_2-normalised representations. All clients start from the same public checkpoint for f_k and the same initialisation of p_k and g_k, and train locally thereafter. Contrastive learning acts in the projection space z, while the encoder space h stays close to the pretrained model and feeds the linear probe, so class summaries are kept in both spaces. Symbols are listed in Table 2.

3.2 Class descriptors

Once per round, client k passes over its data to describe each class in both spaces, encoder and projection, using a sample count, a *centroid*, and a *spread* (Fig. 1(a)). The *centroid* is the class's mean direction. The *spread* is a low-rank Gaussian composed of the top r *eigenvectors* of the class covariance matrix, their *eigenvalues*, and a single *residual variance* for the remaining $d - r$ directions. This lowers communication costs from $O(d^2)$ to $O(rd)$, with *eigenvalues* shrunk toward isotropy to keep small classes well-conditioned. Each part of the spread is used downstream: the *eigenvectors* and *eigenvalues* set the Mahalanobis geometry of the score (Sec. 3.3), while the *residual variance* fixes the score's reference scale and the isotropic part of the Gaussian from which locally absent classes are sampled (Sec. 3.5). Unlike CCVR (Luo et al., 2021), which sends class-conditional Gaussians alongside model weights, our method replaces the weights entirely.

Together, the count, the *centroid* and the *spread* form a *Gaussian component*, and only Gaussian components are exchanged every round. Appendix A.2 gives the estimator, the shrinkage rule, and the reduced descriptors transmitted by classes too small to support a covariance.

3.3 Gaussian component bank

The server collects the components of all clients into a single bank $\mathcal{B}$ (Fig. 1(b)). Components are kept side by side and are never averaged, so the bank records how the same class looks at different sites, and a class held by several clients is represented by several components. Together, the components of one class form a Gaussian mixture over sites; this is the mixture FedGME shares, and the numerator of equation 2 scores a representation against it.

To train against this bank, a client needs a single number saying how well one of its representations fits a given component. Representation norms drift during training and differ across clients, whereas directions remain comparable, which is why every summary and every score is computed on normalised representations. A unit-norm representation u is scored against component j by the log-density of that component's Gaussian,

$$s_j(u) = -\frac{m_j^2(u)}{2T}, \qquad m_j^2(u) = \sigma_{\text{ref}}^2 \, (u - \mu_j)^\top \Sigma_j^{-1} (u - \mu_j), \tag{1}$$

where μ_j is the component's centroid, Σ_j its low-rank covariance, σ_{ref}^2 the median residual variance in the bank and T a temperature.

The score reads as a Mahalanobis distance to the class, so deviations along the directions in which the component is wide are discounted, while deviations along its narrow directions are not: a tight class asks representations to sit close to its centroid, a broad one accepts more variation. The factor σ_{ref}^2 only sets the scale, so that a single T fits the whole bank. Two banks are built, one in z and one in h. Because Σ_j is a rank-r update of a scaled identity, m_j^2 is evaluated in $O(rd)$ and Σ_j is never formed; Appendix A.3 gives the reduced form, the role of σ_{ref}^2, and why the Gaussian normalizing constant is dropped.

3.4 Representation learning objective

Each client receives a bank of components from other clients and then trains against this bank while keeping its own data local. Components from the previous communication round are treated as fixed targets, so gradients are propagated only through the client's current representations. Each window is augmented twice with stochastic, label-preserving transformations standard for EEG (Rommel et al., 2022), such as amplitude scaling, additive Gaussian noise, small time shifts and channel masking; one view serves as the anchor and the other as its positive. I denotes the anchor views of a minibatch. The training objective has three parts with distinct jobs: one term carries information from other clients into the representation, one organizes the client's own batch, and one holds the encoder to the shared pretrained reference.

Cross-client term. This is the only term through which another client's data influences the encoder: it informs a window of where other sites place its class and pulls it toward that location. For $u \in \{z, h\}$, each window is compared with the components that other clients contributed for its class (Fig. 1(c)):

$$\mathcal{L}_{\text{bank}}^u = -\frac{1}{|I_B|} \sum_{i \in I_B} \log \frac{\sum_{j \in P_i} \exp s_j(u_i)}{\sum_{j \in \mathcal{B}^u} \omega_j \exp s_j(u_i)}, \qquad P_i = \{j : c_j = y_i, \; k_j \neq k\} \tag{2}$$

where $I_B = \{i \in I : P_i \neq \emptyset\}$ and $u_i = z_i$ or $\hat{h}_i$. The client's own component is excluded from the numerator: it was estimated from the very windows being scored, so the only way to lower the term is to move toward other clients' components. It remains in the denominator, where it pushes windows away from their own client's centroid and so counteracts collapse. Components of locally absent classes get weight $\omega_j = \omega_{\text{abs}} < 1$ (otherwise 1), limiting negative pressure the client has no evidence for (Li et al., 2021; Zhang et al., 2022). We write $\mathcal{L}_{\text{cross}} = \sum_u \lambda_{\text{bank}}^u \mathcal{L}_{\text{bank}}^u$.

Patient-aware local term. The bank describes each class only at the level of whole clients, and only as of the preceding round, so it cannot prevent a class from collapsing locally and says nothing about how the windows within a batch should relate to one another. We therefore add supervised contrastive and class-conditional spreading terms (Khosla et al., 2020; Chen et al., 2022), made patient-aware (Fig. 1(d)). Patient identity is a strong factor of variation in EEG — subjects can be identified from representations trained for an unrelated task (Wang et al., 2020; Özdenizci et al., 2020), and unaligned embeddings have been reported to cluster by subject rather than by class (Shen et al., 2023) — so a supervised contrastive loss that treats all same-class pairs alike may be dominated by patient-specific structure. The aim is to keep patient variability from defining the class structure, leaving the spread of each component to capture that variability instead. We therefore treat same-class windows differently, depending on whether they come from the same patient. In $\mathcal{L}_{\text{sup}}$ the positives of an anchor are its second view and the same-class windows of *other* patients, which are

up-weighted, while same-class windows of the *same* patient act as negatives and are pushed apart rather than pulled together; in the spreading term $\mathcal{L}^u_{\text{cnce}}$ the only positive is the second view and the other same-class windows are negatives, with cross-patient pairs down-weighted. Both terms and their weights are defined in Appendix A.4, and we write $\mathcal{L}_{\text{local}} = \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \sum_u \lambda^u_{\text{cnce}}\mathcal{L}^u_{\text{cnce}}$.

Pretrained anchoring. Both terms above move the encoder, and neither holds it near the representation every client started from. $\mathcal{L}_{\text{anchor}} = \lambda_{\text{an}}\text{mean}_i\,[1 - \cos(h_i, h_i^0)]$, evaluated on every window of the batch, limits drift from the frozen pretrained encoder $h^0 = f_0(x)$ (Fig. 1(e); Appendix A.5). Because every client is held to the same reference, this is also what keeps a class pointing in nearly the same direction at every client, and so what makes components estimated at one site comparable at another. With all loss weights fixed across runs:

$$\boxed{\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{anchor}}} \tag{3}$$

3.5 Locally absent classes

Every term so far acts on classes the client observes. A locally absent class enters $\mathcal{L}_{\text{rep}}$ only through the denominator of the cross-client term, and cross-entropy on the observed labels keeps pushing its logit down, so the client learns where such a class is *not*, never what it looks like. We treat the two halves of the model separately: the encoder preserves the pretrained geometry of those classes, and the probe is trained on representations sampled from the bank.

Absent-subspace anchoring. A client cannot learn how an unseen class maps into h, but it can avoid overwriting where the shared pretrained encoder already places it. From the consensus centroids of its absent classes in $\mathcal{B}^h$, the client forms an orthonormal basis Q_A of the few directions that carry them, at most $0.1\,d_h$, and penalizes movement away from the pretrained encoder inside that subspace only, leaving every other direction free ($\mathcal{L}_{\text{freeze}}$, Fig. 1(e); Appendix A.6). The penalty applies to every sample, not only to those of a particular class, because the constraint is on the encoder map, rather than on a class. The term is a soft, representation-space analogue of gradient projection (Saha et al., 2021), and is added to the encoder objective as $\mathcal{L}_{\text{rep}} + \lambda_{\text{fr}}\mathcal{L}_{\text{freeze}}$ (Sec. 3.6).

Synthetic replay. Preserving that geometry still leaves the classifier with nothing to separate, since it has never seen an example of the class. However, the components in the bank are Gaussians, so they can be sampled. For every component of an absent class, the client draws feature vectors in h, not raw EEG windows, from that component's Gaussian, widened by a factor ζ, normalizes them and labels them with the component's class (Fig. 1(f); Appendix A.6). These synthetic features are mixed with the features of the client's real windows and classified in a single softmax, so the absent class must be separated from the present classes on real data: its draws must score highest for it, and the real windows of present classes must not. Both inputs are detached, so this trains only g_k: letting the draws reach the encoder would train it to reproduce its own low-rank approximation. We rescale the sampling width ζ during training so that synthetic samples are as spread around their centroid as the client's real windows are around theirs, measured by average cosine similarity. At a fixed width, synthetic samples cluster far more tightly than real data, so the probe fits a small region around the synthetic cloud, and real windows of that class, which spread further from the centroid, fall outside it into the regions of neighbouring present classes.

3.6 Training, communication, and readout

Training has two phases. In the *representation phase*, 25 synchronous rounds of one local epoch each, the encoder and projector minimise $\mathcal{L}_{\text{rep}} + \lambda_{\text{fr}}\mathcal{L}_{\text{freeze}}$, and clients exchange their components with the server after every epoch (the bank exchange). The linear probe g_k is trained alongside only through $\lambda_{\text{replay}}\mathcal{L}_{\text{replay}}$ on detached representations, so no cross-entropy gradient reaches the encoder. In the *probe phase*, 10 further rounds, the encoder and projector are frozen and only g_k is trained; every method is evaluated this way, with the same probe architecture and budget (Appendix B.6).

All clients are drawn from the TUH EEG Corpus (Obeid & Picone, 2016). Each patient is assigned to exactly one client, and label skew is imposed deliberately, by omitting selected classes from some clients in multiclass tasks and by varying class ratios in binary ones (Appendix B.3). Each client is

Table 1: (a) BAcc over all and over locally absent classes (Abs.; averaged over clients with at least one, on average 2.6 per client on TUEV, 1.2 and 0.6 on TUSZ). (b) Effective rank and uniformity (lower is more uniform) of the normalised encoder embeddings on TUAB, and communication per client and round, computed per corpus as uplink plus downlink with every client holding every class (Appendix A.7). Means over clients and five seeds; best in bold; standard deviations in Appendix C.

(a)	TUEV				TUSZ				TUAB	
	BAcc		Abs.		BAcc		Abs.		BAcc	
Method K:	5	10	5	10	5	10	5	10	5	10
Local	0.316	0.300	0.000	0.000	0.284	0.283	0.012	0.009	0.721	0.716
FedAvg	0.348	0.353	0.103	0.104	0.286	0.293	0.054	0.000	0.742	0.737
FedPCL	0.328	0.319	0.000	0.000	0.279	0.282	0.010	0.011	0.742	0.729
FedPCL-S	0.336	0.321	0.000	0.000	0.279	0.278	0.011	0.011	0.748	0.732
FedPCL-SM	0.327	0.317	0.000	0.000	0.280	0.284	0.011	0.011	0.747	0.730
FedGME	**0.408**	**0.377**	**0.164**	**0.183**	**0.339**	**0.350**	**0.150**	**0.295**	**0.749**	**0.740**

(b)	TUAB geometry				Communication ($\times 10^3$ floats)					
	eRank		Unif.		TUEV		TUSZ		TUAB	
Method K:	5	10	5	10	5	10	5	10	5	10
Local	97.6	96.4	−1.90	−1.96	**0**	**0**	**0**	**0**	**0**	**0**
FedAvg	127.0	127.0	−2.91	−3.15	13,082	13,082	13,081	13,081	13,080	13,080
FedPCL	41.3	32.1	−1.01	−0.93	2.3	4.2	1.5	2.8	0.8	1.4
FedPCL-S	142.5	**134.4**	−3.21	−3.19	2.3	4.2	1.5	2.8	0.8	1.4
FedPCL-SM	**139.8**	130.0	**−3.30**	**−3.27**	18.8	34.4	12.5	22.9	6.3	11.5
FedGME	124.2	119.0	−2.94	−2.76	92.8	170.1	61.9	113.4	30.9	56.7

evaluated on a common test set that includes its locally absent classes. Results are averaged over five seeds, each fixing patient allocation, class omissions and initialisation, so the reported variance reflects partition heterogeneity together with training noise. All methods share one optimisation setup (AdamW, batch size 64, class-balanced sampling; Appendix B.5) and differ only in their loss terms.

4 Results

We evaluate FedGME against five baselines. **Local** trains each client independently without collaboration, and **FedAvg** exchanges and aggregates model parameters each round. The remaining three are ablations of FedGME, each isolating one design decision. **FedPCL** communicates one normalised centroid per observed class; for comparability with our arms, we train the encoder locally rather than keeping it frozen. **FedPCL-S** keeps that payload and substitutes the patient-aware contrastive loss we adopt (Sec. 3.4), testing whether the loss alone mitigates representation collapse. **FedPCL-SM** additionally replaces prototypes with per-client Gaussian components and confines the contrastive loss to each client's own batch, so the component bank becomes the sole carrier of cross-client information, while its generative capacity is left unused. **FedGME** exploits it, using the components both to protect the directions of classes a client never observes and to sample features for them. We evaluate across three datasets from the TUH EEG Corpus (Obeid & Picone, 2016) (Appendix B.1), covering EEG event classification (TUEV, six classes), seizure-type classification (TUSZ, four classes) and normal/abnormal classification (TUAB, binary) at $K = 5$ and $K = 10$. Table 1 reports balanced accuracy over all and over locally absent classes, communication overhead, and, on TUAB, effective rank (Roy & Vetterli, 2007; Garrido et al., 2023) and uniformity (Wang & Isola, 2020) of the normalised encoder representations $\hat{h}$; standard deviations, corpus-specific metrics and the geometry of all corpora are in Appendix C.

FedGME consistently attains the highest balanced accuracy on both multiclass corpora, TUEV and TUSZ (Table 1). Balanced accuracy is the critical metric here: some classes, such as absence seizures in TUSZ (0.35% of training windows), are extremely rare, so aggregate metrics can mask poor minority-class performance. On TUEV at $K = 5$, FedGME reaches a balanced accuracy of 0.408, versus 0.348 for FedAvg and 0.316 for Local. On TUSZ the gain is more consistent across federation sizes: FedGME reaches 0.339 at $K{=}5$ and 0.350 at $K{=}10$, exceeding FedAvg by

0.053 and 0.057, and both gaps exceed four pooled standard deviations of the seed variability. The remaining baselines reveal distinct, corpus-dependent limitations. On TUEV, the prototype-based methods improve over Local by only 0.011 to 0.021 absolute balanced accuracy, and on TUSZ they offer no advantage at all, staying within 0.005 of Local. FedAvg does better on TUEV ($+0.032$ and $+0.053$ over Local), but despite transmitting the full model gains only 0.002–0.010 on TUSZ. Corpus-specific metrics (Yang et al., 2023) agree with balanced accuracy (Appendix C.1) except on TUEV at $K = 10$.

Restricting balanced accuracy to the classes a client never trains on (Abs.) isolates the setting FedGME targets. Local and the prototype-based methods stay at or near zero (≤ 0.012). The prototype-based methods do receive other clients' prototypes of these classes, but use them only as targets for the encoder; the linear probe is trained on local labels alone, never sees an example of an absent class, and learns to suppress it. FedAvg recovers part of TUEV (0.103 and 0.104), plausibly because its shared encoder sees every class through other clients' updates, but little of TUSZ (0.054 and 0.000). FedGME reaches 0.164 and 0.183 on TUEV and 0.150 and 0.295 on TUSZ, 1.6–$2.8\times$ FedAvg wherever FedAvg is non-zero. FedPCL-SM, whose bank contains components of the absent classes but uses them only in the denominator of the bank term, likewise stays near zero: the bank's value lies in how it is used, not what it stores.

FedGME does not win everywhere. On TUEV at $K = 10$, FedAvg outperforms it on $F1_w$ (0.636 vs. 0.596) and Cohen's κ (0.351 vs. 0.267; Table 7), although FedGME keeps the higher balanced accuracy. Both metrics are dominated by the majority BCKG class (67% of the TUEV test set), so this is consistent with FedGME trading some majority-class accuracy for recall on rare and absent classes. Increasing K from 5 to 10 also halves the number of patients per client, and FedGME's TUEV balanced accuracy drops from 0.408 to 0.377, plausibly because each component is then estimated from fewer patients, whereas on TUSZ it remains stable within one standard deviation.

On TUAB, all methods cluster tightly: balanced accuracy differs by at most 0.028 across the six methods, every federated method improves over Local, and FedGME edges out FedAvg by 0.007 at $K = 5$ and 0.003 at $K = 10$. This is expected: TUAB is balanced and has no locally absent classes, so none of the structural challenges FedGME targets are present here, which makes it a control. What TUAB does expose is a difference in representation quality, which we turn to next.

Lacking any collapse-prevention mechanism, FedPCL exhibits typical supervised-contrastive behaviour: it collapses within class, giving the lowest effective rank and the worst uniformity on TUAB (41.3 and -1.01 at K=5) and on TUEV. On TUAB this collapse costs nothing—FedPCL matches FedAvg at 0.742 balanced accuracy, since binary probing only needs one discriminative direction. Yet the collapse itself is removed by the adopted patient-aware loss rather than by the component bank: with the same prototype payload, FedPCL-S raises TUAB's effective rank to 142.5 and 134.4 (up to $4.2\times$) and improves uniformity to -3.21 and -3.19. However, avoiding collapse alone does not guarantee downstream gains. On TUEV, FedPCL-S attains the highest effective rank and the lowest uniformity at $K = 5$, yet it outperforms FedPCL by only 0.008 balanced accuracy, and on TUSZ FedAvg attains the highest effective rank yet trails FedGME by 0.053–0.057 (Table 11). Taken together, both metrics detect collapse but do not predict accuracy.

These gains come at a fraction of the communication cost. Counting uplink and downlink for every method, FedAvg transmits about 13.1M floats per client per round, while FedGME transmits between 30.9K (TUAB, $K = 5$) and 170.1K (TUEV, $K = 10$)—0.2% to 1.3% of FedAvg's payload — and no model weights leave a site. This overhead grows linearly with K and the number of classes, as components are not merged, but is independent of model size (Appendix A.7).

5 Discussion

The central finding of this study is that exchanging class distributions improves performance specifically on rare, unevenly distributed classes, while offering little benefit on balanced, abundant data. On TUEV and TUSZ, where rare events and seizure types matter clinically, our approach yields consistent balanced-accuracy gains—raising performance on locally absent classes to 0.15–0.30 where no baseline exceeds 0.104 (Table 1)—whereas differences across methods nearly vanish on the balanced TUAB dataset. The ablation ladder illustrates which mechanisms drive these improvements. Moving from a single centroid (FedPCL) to the patient-aware contrastive loss (FedPCL-S, Sec. 3.4)

reduces representation collapse and yields modest accuracy gains on TUEV. However, replacing the centroid with a Gaussian component scored against the bank (FedPCL-SM) adds no gain on the multiclass corpora: the bank does contain the absent classes, but without the absent-class mechanisms they enter training only as negatives, and the linear probe, trained on local labels alone, never sees an example of them. Substantial gains appear only in FedGME, which uses the components to anchor and sample those classes. Since the last two steps of the ladder each change more than one element (Appendix B.4), we conclude that the value of the component bank stems from the mechanisms that exploit it rather than from the statistical richness of the stored data alone.

These gains highlight a limitation of standard federated approaches such as FedAvg. Although FedAvg exchanges far more model information, parameter averaging under non-IID label distributions is known to bias the model toward majority classes (Zhao et al., 2018; Zhang et al., 2022), which plausibly outvotes rare classes regardless of local representation quality; its low accuracy on locally absent classes (at most 0.104) is consistent with this. FedGME avoids this limitation by preserving individual client contributions: every site holding a class contributes its own component without averaging, so rare classes remain represented in the bank.

FedGME is not inherently tied to EEG, as both components are modality-agnostic. The component bank needs only a pretrained encoder, and its cost scales with d_h and r rather than with model size (Appendix A.7), so it applies to other biosignals such as ECG, where rare rhythms can likewise be concentrated at few hospitals, and to vision, where label skew with missing classes is widely studied (Luo et al., 2021; Zhang et al., 2022) and strong public backbones can serve as the shared frozen reference. We expect the benefit wherever sites lack classes that others hold, and little on balanced tasks, as on TUAB. The patient-aware loss transfers wherever one grouping factor dominates the representation, such as the subject in ECG or the patient or scanner in medical imaging; without such a factor, it reduces to the losses it builds on. Whether one low-rank Gaussian per client and class suffices for higher-dimensional vision features remains untested.

Finally, these findings must be interpreted within the scope of our experimental setup. All clients are simulated from a single archive (TUH), so patient-disjoint partitions reproduce label skew and inter-patient variability but not the hardware or preprocessing differences between institutions, which could challenge the shared pretrained initialisation. All experiments use one backbone (BIOT), and we do not study sensitivity to the covariance rank r, the absent-class weight ω_{abs} or the sampling width ζ. Our partitions place every class at several sites, so a class held by a single site remains untested. Similarly, a class with few windows at a site falls back to a centroid-only or isotropic descriptor (Appendix A.2), so the rarest classes, such as TUSZ absence seizures (60 training windows in total), can be shared with less structure than the method is designed to exploit. Moreover, we deliberately separate encoder and classifier training (Sec. 3.6) so that representation geometry is measured without confounding cross-entropy, making end-to-end training a natural next step. Furthermore, because components are preserved rather than merged, communication grows linearly with the number of clients, suggesting aggregation or clustering at larger scales. Lastly, no model weights leave a site, but the exchanged components are not private by construction: a rare class at one site may be summarised from only one or two patients, and EEG representations are known to encode patient identity (Özdenizci et al., 2020). Formal guarantees, such as differential privacy on the transmitted statistics, remain future work.

6 Conclusion

We introduced FedGME, a federated learning method for clinical EEG that communicates one low-rank Gaussian component per client and class instead of model parameters or single-point prototypes. Clients train against a shared bank of these components, which aligns representations across sites and lets each client learn classes it never observes locally. On the two multiclass corpora, FedGME recognises such classes and improves balanced accuracy over FedAvg at 0.2%–1.3% of its payload. The ablations indicate that this gain comes from the mechanisms that use the bank for absent classes rather than from storing richer statistics alone. On the balanced binary task, where no class is absent, FedGME performs on par with FedAvg. Future work includes end-to-end training of encoder and classifier, evaluation on data from multiple institutions, formal privacy guarantees, and merging components to bound communication as the number of sites grows.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods through code generation and with summarizing results. We have not used generative AI tools for other tasks requiring disclosure, including developing theoretical or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, designing research methodology or experiments, qualitative or thematic data analysis, or interpreting results; other required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools to assist with scientific writing, including improving the clarity, wording, and presentation of the manuscript. We reviewed all AI-assisted work. AI-generated code was manually inspected and tested for correctness before use, and AI-assisted text was reviewed and revised by the authors before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the TUH EEG Corpus (Obeid & Picone, 2016), a publicly available, de-identified dataset released for research use under its established data use terms; we do not collect, process, or have access to any personally identifiable patient information, and no human-subject experimentation was conducted as part of this study. We follow the corpus's intended usage conditions and standard practices for EEG data processing and evaluation.

REPRODUCIBILITY STATEMENT

We have aimed to make every component of FedGME reproducible from the paper and its appendix. Sec. 3 describes the method, and Appendix A gives its full specification: the estimation of the class descriptors (Appendix A.2), the bank score (Appendix A.3), all loss terms (Appendices A.4–A.6) and the communication accounting (Appendix A.7). Appendix B describes the datasets and their splits, the pretrained checkpoint, the construction of client partitions, the baselines and the evaluation protocol. All corpora are publicly available from the TUH EEG archive (Obeid & Picone, 2016), and the encoder is initialised from a public BIOT checkpoint (Yang et al., 2023). All results are averaged over five seeds, and standard deviations are reported in Appendix C. The source code, including the configuration files with all hyperparameters and the seeds used to generate the client partitions, will be made publicly available after the review process.

REFERENCES

Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 232–241. PMLR, 2019.

V. S. Arulmurugan, R. Aarthy, S. Sesha Vidhya, and Saranya K. Federated learning-enabled privacy-preserving framework for seizure forecasting and affective state analysis using multi-modal EEG-ECG data. *Scientific Reports*, 16:20382, 2026. doi: 10.1038/s41598-026-50003-5.

Saleh Baghersalimi, Tomas Teijeiro, David Atienza, and Amir Aminifar. Personalized real-time federated learning for epileptic seizure detection. *IEEE Journal of Biomedical and Health Informatics*, 26(2):898–909, 2022. doi: 10.1109/JBHI.2021.3096127.

Saleh Baghersalimi, Tomas Teijeiro, Amir Aminifar, and David Atienza. Decentralized federated learning for epileptic seizures detection in low-power wearable systems. *IEEE Transactions on Mobile Computing*, 23(5):6392–6407, 2024. doi: 10.1109/TMC.2023.3320862.

Calvin Chan, Qingqing Zheng, Chengjian Xu, Qiong Wang, and Pheng-Ann Heng. Adaptive federated learning for EEG emotion recognition. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024. doi: 10.1109/IJCNN60899.2024.10650004.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Re. Perfectly balanced: Improving transfer and robustness of supervised contrastive

learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3090–3122. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/chen22d.html.

Mingqing Chen, Ananda Theertha Suresh, Rajiv Mathews, Adeline Wong, Cyril Allauzen, Françoise Beaufays, and Michael Riley. Federated learning of n-gram language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 121–130. Association for Computational Linguistics, 2019. doi: 10.18653/v1/K19-1012.

Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, Safaa Abdel-Razek, and Chuan Liu. Fed-ESD: Federated learning for efficient epileptic seizure detection in the fog-assisted internet of medical things. *Information Sciences*, 630:403–419, 2023. doi: 10.1016/j.ins.2023.02.052.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Tracy Glauser, Elinor Ben-Menachem, Blaise Bourgeois, Avital Cnaan, Carlos Guerreiro, Reetta Kälviäinen, Richard Mattson, Jacqueline A. French, Emilio Perucca, and Torbjorn Tomson. Updated ILAE evidence review of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. *Epilepsia*, 54(3):551–563, 2013. doi: 10.1111/epi.12074.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3821–3830. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/graf21a.html.

Kuangpu Guo, Yuhe Ding, Jian Liang, Zilei Wang, Ran He, and Tieniu Tan. Exploring vacant classes in label-skewed federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16960–16968, 2025. doi: 10.1609/aaai.v39i16.33864.

Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large CNNs at the edge. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Chenghao Huang, Xiaolu Chen, Yanru Zhang, and Hao Wang. FedCoSR: Personalized federated learning with contrastive shareable representations for label heterogeneity in non-iid data. *IEEE Transactions on Cybernetics*, 55(12):5841–5854, 2025. doi: 10.1109/tcyb.2025.3599370. URL https://doi.org/10.1109/tcyb.2025.3599370.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. In *NeurIPS 2018 Workshop on Machine Learning on the Phone and other Consumer Devices*, 2018.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

Xiao Jiang, Gui-Bin Bian, and Zean Tian. Removal of artifacts from EEG signals: A review. *Sensors*, 19(5):987, 2019. doi: 10.3390/s19050987.

Ce Ju, Dashan Gao, Ravikiran Mane, Ben Tan, Yang Liu, and Cuntai Guan. Federated transfer learning for EEG signal classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3040–3045, 2020. doi: 10.1109/EMBC44109.2020.9175344.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10708–10717, 2021. doi: 10.1109/CVPR46437.2021.01057.

Zheng You Lim, Ying Han Pang, Shih Yin Ooi, Sarmela Raja Sekaran, and Yee Jian Chew. FCEEG: Federated learning-based seizure diagnosis through electroencephalogram (EEG) analysis. *Cogent Engineering*, 12(1):2547636, 2025. doi: 10.1080/23311916.2025.2547636.

Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. FedVision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13172–13179, 2020.

Jionghao Lou, Jian Zhang, Zhongmei Li, Lanlan Chen, and Enbo Feng. FedL2T: Personalized federated learning with two-teacher distillation for seizure prediction, 2025.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-IID data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *PMLR*, pp. 1273–1282, 2017.

Anh Nguyen, Tuong Do, Minh Tran, Binh X. Nguyen, Chien Duong, Tu Phan, Erman Tjiputra, and Quang D. Tran. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1824–1830, 2022. doi: 10.1109/IV51971.2022.9827020.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196.

Ozan Özdenizci, Ye Wang, Toshiaki Koike-Akino, and Deniz Erdoğmuş. Learning invariant representations from EEG via adversarial inference. *IEEE Access*, 8:27074–27085, 2020. doi: 10.1109/ACCESS.2020.2971600.

Nan Pu, Wenjing Li, Xingyuan Ji, Yalan Qin, Nicu Sebe, and Zhun Zhong. Federated generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28741–28750, 2024.

Yu Qiao, Md. Shirajum Munir, Apurba Adhikary, Huy Q. Le, Avi Deb Raha, Chaoning Zhang, and Choong Seon Hong. MP-FedCL: Multi-prototype federated contrastive learning for edge intelligence. *IEEE Internet of Things Journal*, 11(5):8604–8623, 2024. doi: 10.1109/JIOT.2023.3320546.

Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41473–41497. PMLR, 2024.

Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on EEG: a systematic comparison. *Journal of Neural Engineering*, 19 (6):066020, 11 2022. doi: 10.1088/1741-2552/aca220. URL https://doi.org/10.1088/1741-2552/aca220.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610. IEEE, 2007.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019. doi: 10.1088/1741-2552/ab260c.

Raghdah Saemaldahr and Mohammad Ilyas. Patient-specific preictal pattern-aware epileptic seizure prediction with federated learning. *Sensors*, 23(14):6578, 2023. doi: 10.3390/s23146578.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3AOj0RCNC2.

Seonguk Seo, Jinkyu Kim, Geeho Kim, and Bohyung Han. Relaxed contrastive learning for federated learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12279–12288, 2024. doi: 10.1109/CVPR52733.2024.01167.

Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018. doi: 10.3389/fninf.2018.00083.

Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2496–2511, 2023. doi: 10.1109/TAFFC.2022.3164516.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pp. 4077–4087, 2017.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022a.

Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 19332–19344, 2022b.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025.

Min Wang, Jiankun Hu, and Hussein A. Abbass. BrainPrint: EEG biometric identification based on analyzing brain connectivity graphs. *Pattern Recognition*, 105:107381, 2020. doi: 10.1016/j.patcog.2020.107381.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=c2LZyTyddi.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26311–26329. PMLR, 2022.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated Learning with Non-IID Data. *arXiv preprint arXiv:1806.00582*, 2018. doi: 10.48550/arxiv.1806.00582. URL https://arxiv.org/abs/1806.00582.

Yanbing Zhou, Xiangmou Qu, Chenlong You, Jiyang Zhou, Jingyue Tang, Xin Zheng, Chunmao Cai, and Yingbo Wu. FedSA: a unified representation learning via semantic anchors for prototype-based federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21): 23009–23017, 2025. doi: 10.1609/aaai.v39i21.34464. URL https://doi.org/10.1609/aaai.v39i21.34464.

A Appendix

Appendix A gives the definitions, estimation procedures and derivations behind Sec. 3, one subsection per main-text subsection; Appendix B describes the experimental setup and Appendix C reports additional results.

A.1 Notation

Table 2: Symbols used in Sec. 3 and Appendix A.

Symbol	Meaning
K, k	number of clients; index of a client
$\mathcal{C}$, $\mathcal{C}_k$	full label set; classes observed at client k
$\mathcal{C} \setminus \mathcal{C}_k$	classes locally absent at client k
x, y_i, $\mathrm{pt}(i)$	input window; its label; its patient identifier (never transmitted)
$f_k,\ p_k,\ g_k$	encoder, projector, linear probe of client k
f_0, h^0	frozen pretrained encoder and its output
$h \in \mathbb{R}^{d_h}$, $z \in \mathbb{S}^{d_z-1}$	encoder and (ℓ_2-normalised) projection representations; $d_h = 256$, $d_z = 64$
$\hat{h}$, $\hat{h}^0$	ℓ_2-normalised encoder representation and its pretrained counterpart
u	a normalised representation in either space, $u \in \{z, \hat{h}\}$
I, I_B	anchor views in a minibatch; those with at least one remote positive
i^+	the augmented view paired with anchor i
$P(i)$, $N(i)$	positives and negatives of anchor i in the local term
ν_{ip}, β_{ij}	positive weights in $\mathcal{L}_{\mathrm{sup}}$; negative weights in $\mathcal{L}^u_{\mathrm{cnce}}$
η, γ	cross-patient attraction gain (≥ 1); within-batch cross-patient repulsion weight ($\in [0,1]$)
B_{kc}, $\bar{h}_{kc}$	windows of class c in the batch; their normalised mean $\hat{h}$
n_{kc}	number of windows of class c at client k
μ_{kc}	class centroid of client k
$U_j \in \mathbb{R}^{d\times r}$, Λ_j	top-r eigenvectors and eigenvalues of component j
σ^2_j, σ^2_{ref}	residual variance of component j; median residual variance in the bank
r	rank of the retained covariance ($r = 7$)
$\mathcal{B}$, $\mathcal{B}^z$, $\mathcal{B}^h$	Gaussian component bank, in each space
j, c_j, k_j	a component; its class; the client that contributed it
P_i	components of class y_i contributed by other clients
$s_j(u)$, $m^2_j(u)$	score of u under component j; its Mahalanobis distance
T, τ, τ_s	temperatures of the bank score, supervised, and spreading terms
ω_j, ω_{abs}	component weight in the bank denominator; its value for absent classes
Q_A	orthonormal basis of the absent-class subspace in h
ζ, θ	width factor applied when sampling synthetic features; EMA coefficient of its calibration
n_{rep}	synthetic draws per absent component and step
$\lambda_\bullet$	loss weights

A.2 Class descriptors: estimation

Estimation. All statistics are computed on ℓ_2-normalised representations (Sec. 3.1). For class c at client k with n_{kc} windows in space u, we take the centroid $\mu_{kc} = \overline{u}/\|\overline{u}\|_2$ and the top-r eigenpairs (U, Λ) of the centred second-moment matrix. The matrix itself is never formed: the eigenpairs are obtained by randomised subspace iteration on the centred data in $O(n_{kc}dr)$ time and $O(dr)$ memory,

and the residual variance follows from the trace,

$$\sigma^2 = \frac{1}{d-r}\Big(\tfrac{1}{n_{kc}}\textstyle\sum_i \|u_i - \bar{u}\|_2^2 - \sum_{l=1}^{r}\lambda_l\Big), \tag{4}$$

which needs one $O(n_{kc}d)$ pass. The component is therefore the spiked covariance

$$\Sigma = U\Lambda U^\top + \sigma^2\big(I - UU^\top\big), \tag{5}$$

i.e. r estimated directions and an isotropic remainder. This is the Σ_j that enters the score of equation 1.

Shrinkage. Eigenvalues estimated from few windows are biased outward, and a class with $n_{kc} \approx r$ yields a near-singular Σ that dominates any softmax it enters — a single under-supported class at one client would otherwise capture the bank denominator for every other client. We therefore shrink toward the isotropic target $\bar{\sigma}^2 = \mathrm{tr}(\Sigma)/d$,

$$\tilde{\lambda}_l = (1-\alpha)\lambda_l + \alpha\bar{\sigma}^2, \qquad \tilde{\sigma}^2 = (1-\alpha)\sigma^2 + \alpha\bar{\sigma}^2, \qquad \alpha = 0.2, \tag{6}$$

with α a fixed hyperparameter; support-dependence is handled by the tiered descriptors below rather than by α. Shrinkage is applied before transmission, so the bank contains only well-conditioned components and no client has to re-condition what it receives. An absolute and a relative floor are applied to the residual variance so that no component can become numerically singular.

Reduced descriptors for small classes. Below a certain support, an estimated covariance carries no information worth transmitting, so the descriptor degrades gracefully instead. A class with $n_{kc} < n_{\mathrm{cent}}$ transmits its centroid alone and is scored with $\Sigma = \sigma_{\mathrm{ref}}^2 I$, which reduces m_j^2 to $\|u - \mu_j\|_2^2 = 2(1 - u^\top\mu_j)$, i.e. a cosine distance. A class with $n_{\mathrm{cent}} \le n_{kc} < n_{\mathrm{shape}}$ additionally transmits its eigenvalues, of which the bank retains only $\mathrm{tr}(\Sigma)$ and scores the component as isotropic with $\sigma^2 = \mathrm{tr}(\Sigma)/d$. Eigenvectors are transmitted only above the second threshold. The three cases share one scoring routine, since the first two are equation 5 with $r = 0$.

A.3 Scoring a representation against a component

Scoring in $O(rd)$. Because equation 5 is a rank-r update of a scaled identity, its inverse has the same form,

$$\Sigma^{-1} = U\Lambda^{-1}U^\top + \sigma^{-2}\big(I - UU^\top\big), \tag{7}$$

and with $\delta = u - \mu_j$ and $a = U_j^\top\delta \in \mathbb{R}^r$ the quadratic form of equation 1 becomes

$$m_j^2(u) = \sigma_{\mathrm{ref}}^2\left[\sum_{l=1}^{r}\frac{a_l^2}{\lambda_{jl}} + \frac{\|\delta\|_2^2 - \|a\|_2^2}{\sigma_j^2}\right]. \tag{8}$$

The first term measures deviation inside the component's principal subspace, scaled by how wide the class is in each direction; the second measures the part of δ orthogonal to that subspace against a single residual scale. Evaluating equation 8 costs $O(rd)$ for a and $O(d)$ for $\|\delta\|_2^2$, and Σ_j and Σ_j^{-1} are never instantiated. Scoring a batch of B representations against all $|\mathcal{B}|$ components costs $O(B|\mathcal{B}|rd)$, the same order as the dense inner products of an ordinary contrastive loss up to the factor r.

The reference scale. With the factor σ_{ref}^2, a component of typical width scores like a cosine similarity, so a single temperature T fits the whole bank; without it, scores run orders of magnitude larger and the softmax over components saturates. The factor is common to every component, so it changes no ranking — it only places the scores on a scale a single T can serve.

Omitted normalising constant. The exact Gaussian log-density adds $-\frac{1}{2}\big(\sum_l \log\lambda_{jl} + (d-r)\log\sigma_j^2\big)$ to equation 1. We omit it. Retaining it would add a per-component bias proportional to $-\log|\Sigma_j|$ inside the softmax over components, so a client whose class happens to be tight would raise the score of its own component for every representation irrespective of distance, and clients would compete on estimated class volume rather than on agreement. Dropping the term makes s_j a width-normalised distance rather than a calibrated density, which is what the loss requires.

A.4 PATIENT-AWARE LOCAL TERM

Both local terms of Sec. 3.4 are contrastive losses over a minibatch with anchor views I, each anchor i paired with a second view i^+:

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|I|}\sum_{i\in I}\frac{1}{\nu_i}\sum_{p\in P(i)}\nu_{ip}\log\frac{\exp(z_i^\top z_p/\tau)}{\sum_{j\neq i}\exp(z_i^\top z_j/\tau)}, \tag{9}$$

$$\mathcal{L}^u_{\text{cnce}} = -\frac{1}{|I|}\sum_{i\in I}\log\frac{\exp(u_i^\top u_{i^+}/\tau_s)}{\sum_{j\in N(i)}\beta_{ij}\exp(u_i^\top u_j/\tau_s)}. \tag{10}$$

In equation 9 the positives are the second view and the same-class windows of *other* patients, $P(i) = \{i^+\}\cup\{p : y_p = y_i,\ \text{pt}(p)\neq\text{pt}(i)\}$, weighted by $\nu_{ii^+} = 1$ and $\nu_{ip} = \eta \geq 1$ otherwise, with $\nu_i = \sum_{p\in P(i)}\nu_{ip}$; all other windows of the batch are negatives, so same-patient samples are pushed apart rather than pulled together. This is the one difference from the supervised contrastive loss of Khosla et al. (2020), which places them among the positives. In equation 10 the only positive is the second view and the negatives are the same-class windows, $N(i) = \{i^+\}\cup\{j\neq i : y_j = y_i\}$, weighted by $\beta_{ij} = 1$ within a patient and $\beta_{ij} = \gamma \in [0, 1]$ across patients, so that the spreading pressure of Chen et al. (2022) acts mainly inside a patient and does not undo the cross-patient attraction of equation 9.

A.5 PRETRAINED ANCHORING

$\mathcal{L}_{\text{anchor}} = \lambda_{\text{an}}\,\text{mean}_i\,[1-\cos(h_i, h_i^0)]$ is evaluated on every window of the batch, not per class, with $h^0 = f_0(x)$ the output of the frozen pretrained encoder. Among the methods we compare, only FedGME uses it (Appendix B.4).

A.6 LOCALLY ABSENT CLASSES

The two mechanisms of Sec. 3.5 are specified below; their effect on the classes they target is reported in the Abs. columns of Table 1 and, with standard deviations, in Table 10.

Absent subspace. Let $A = \mathcal{C}\setminus\mathcal{C}_k$. For each $c\in A$ the client forms the consensus centroid $\bar{\mu}_c^h \propto \sum_{j:c_j=c} w_j\mu_j$, $w_j = n_j/(n_j+n_0)$, over the components in $\mathcal{B}^h$, stacks them into $M_A\in\mathbb{R}^{d_h\times|A|}$, and takes as Q_A the left singular vectors of M_A with $\varsigma_i > \epsilon_A\varsigma_1$, truncated to at most $\lfloor 0.1\,d_h\rfloor$ directions. The truncation matters: $|A|$ centroids span at most $|A|$ directions, but near-collinear centroids would otherwise contribute directions estimated from noise, and pinning those would restrict the encoder without protecting anything. The penalty is

$$\mathcal{L}_{\text{freeze}} = \frac{1}{|I|}\sum_{i\in I}\big\|Q_A^\top(\hat{h}_i - \hat{h}_i^0)\big\|_2^2, \tag{11}$$

evaluated on all windows, because the constraint is on the encoder map rather than on a class: the directions carrying absent classes must keep their pretrained meaning whatever the input. Every client computes Q_A from the same consensus centroids against the same frozen reference, which keeps a centroid reported by one client interpretable inside another client's encoder.

Synthetic replay. For a component j of an absent class, samples are drawn from equation 5 with the covariance widened by ζ,

$$v = \mu_j + \zeta\Big(U_j\Lambda_j^{1/2}\varepsilon + \sigma_j\big(I - U_jU_j^\top\big)\xi\Big), \qquad \varepsilon\sim\mathcal{N}(0, I_r),\ \xi\sim\mathcal{N}(0, I_{d_h}), \qquad \tilde{v} = \frac{v}{\|v\|_2}, \tag{12}$$

which reproduces the component's covariance exactly at $\zeta = 1$ while costing $O(rd_h)$ per draw. We draw n_{rep} samples per absent component per step, concatenate them with the client's own detached batch representations, and pass both through g_k in a single softmax,

$$\mathcal{L}_{\text{replay}} = \text{CE}\Big(\big\{g_k(\text{sg}[\hat{h}_i]), y_i\big\}_{i\in I}\cup\big\{g_k(\tilde{v}), c_j\big\}\Big), \tag{13}$$

where $\text{sg}[\cdot]$ is the stop-gradient. Because every input is a constant, the term trains g_k only; routing the draws into the encoder would train it to reproduce its own low-rank approximation of a distribution

it has never observed. Sharing one softmax with the real windows forces an absent class to compete against the present ones on real data, rather than being learned on a separate, trivially separable problem. In the probe phase (Sec. 3.6), the draws come from the bank of the last representation round.

Width calibration. A component's eigenvalues describe the remote client's class, but the draws also pass through normalisation, and at $\zeta = 1$ the resulting cloud is concentrated near μ_j relative to real data, so the probe carves out a decision region that real windows never reach. We therefore match the mean cosine of the draws to their centroid, $\bar{\psi}_{\text{syn}} = \text{mean}_{j,\tilde{v}} \cos(\tilde{v}, \mu_j)$, to the same quantity on the client's present classes, $\bar{\psi}_{\text{pres}} = \text{mean}_{c \in \mathcal{C}_k} \text{mean}_{i \in B_{kc}} \cos(\hat{h}_i, \bar{h}_{kc})$, where $\bar{h}_{kc}$ is the normalised mean of $\hat{h}_i$ over the windows of class c in the current batch. Writing $v = \mu_j + \zeta d$ in equation 12, $\cos(\tilde{v}, \mu_j) \approx (1 + \zeta^2 \|d\|_2^2)^{-1/2}$. Inverting this for the width that would reproduce $\bar{\psi}_{\text{pres}}$, then smoothing with EMA coefficient θ and clipping, gives

$$\zeta \leftarrow \text{clip}\Big((1-\theta)\,\zeta + \theta\,\zeta \sqrt{\tfrac{\bar{\psi}_{\text{pres}}^{-2} - 1}{\bar{\psi}_{\text{syn}}^{-2} - 1}},\ \zeta_{\min},\ \zeta_{\max} \Big), \tag{14}$$

so that a cloud tighter than real data ($\bar{\psi}_{\text{syn}} > \bar{\psi}_{\text{pres}}$) widens and a cloud that is too broad contracts.

A.7 Communication accounting

A Gaussian component in a space of dimension d costs

$$c(d) = \underbrace{1}_{\text{count}} + \underbrace{d}_{\text{centroid}} + \underbrace{rd}_{\text{eigenvectors}} + \underbrace{r}_{\text{eigenvalues}} + \underbrace{1}_{\text{residual variance}} = (r+1)d + r + 2, \tag{15}$$

and a fully supported client–class summary holds one component in each space, $c(d_z) + c(d_h) = 521 + 2{,}057 = 2{,}578$ scalars for $d_z = 64$, $d_h = 256$ and $r = 7$, independently of the number of model parameters. Classes below the thresholds of Appendix A.2 cost $d + 1$ or $d + r + 2$ per space instead. Each round, a client uploads the summaries of its classes and downloads the bank, so its traffic is

$$\underbrace{\textstyle\sum_{c \in \mathcal{C}_k} \big(c(d_z) + c(d_h)\big)}_{\text{uplink}} + \underbrace{|\mathcal{B}| \big(c(d_z) + c(d_h)\big)}_{\text{downlink}} \;\le\; (K+1)\,|\mathcal{C}|\,\big(c(d_z) + c(d_h)\big), \tag{16}$$

with equality when every client holds every class and all classes are fully supported ($\mathcal{C}_k = \mathcal{C}$, $|\mathcal{B}| = K|\mathcal{C}|$). Table 1 reports this upper bound, which depends only on K and the number of classes, and grows linearly in both because components are never merged. The baselines are counted the same way, uplink plus downlink: FedPCL and FedPCL-S exchange one d_z-dimensional centroid per class, $(K+1)|\mathcal{C}|\,d_z$; FedPCL-SM exchanges z-space components only, $(K+1)|\mathcal{C}|\,c(d_z)$; and FedAvg uploads and downloads all $6{,}540{,}930$ trainable parameters on TUEV (the linear head changes this by at most $1{,}028$ on the other corpora), independently of K.

B Experimental setup

B.1 Datasets

We evaluate on three corpora from the Temple University Hospital EEG archive (Obeid & Picone, 2016), chosen so that the label sets differ in size and in how unevenly they are distributed: a binary recording-level task (TUAB), a four-class task with two very rare classes (TUSZ), and a six-class event-classification task (TUEV). Tables 3–5 give the composition of each. All splits are patient-disjoint, so no patient contributes windows to more than one split, and the same holds of the client partitions built from them (Appendix B.3).

B.2 Pretrained encoder

All clients and all methods, baselines included, initialise f_k from the public BIOT checkpoint `EEG-PREST-16-channels` (Yang et al., 2023), pretrained without labels on about five million 10 s EEG windows from PREST, a private clinical EEG dataset recorded in a hospital sleep lab.

Table 3: Composition of the TUH Abnormal EEG Corpus (TUAB) (Obeid & Picone, 2016). Record-level labels are inherited by all 10 s windows, introducing weak supervision. EEGs are converted to the 16-channel bipolar TCP montage. Validation is a patient-disjoint 80/20 split of the official training set.

Label	Class	Description	Train	Val	Test
0	Normal	No abnormality reported in the clinical EEG interpretation.	147,300 (49.51%)	36,409 (48.53%)	19,907 (53.88%)
1	Abnormal	Any EEG judged abnormal, irrespective of the underlying diagnosis.	150,187 (50.49%)	38,614 (51.47%)	17,038 (46.12%)
Total windows			297,487	75,023	36,945
Total recordings			2,717		276

Table 4: Composition of the TUH Seizure Corpus (TUSZ) (Shah et al., 2018; Obeid & Picone, 2016) using four seizure-type groups. Samples are 10 s windows extracted from annotated seizures; splits are patient-disjoint. Dev serves as the validation split and Eval as the test split; note the shift in class composition between the two.

Label	Class	Description	Train	Dev	Eval
0	CF	Combined focal seizures, merging FNSZ, SPSZ, and CPSZ.	12,373 (71.54%)	3,175 (45.44%)	1,797 (63.95%)
1	GN	Generalized seizures without sufficient evidence for a specific subtype.	4,613 (26.67%)	3,762 (53.84%)	915 (32.56%)
2	AB	Absence seizures with brief generalized spike-and-wave activity.	60 (0.35%)	15 (0.21%)	32 (1.14%)
3	CT	Combined tonic seizures, merging TNSZ and TCSZ.	250 (1.45%)	35 (0.50%)	66 (2.35%)
Total windows			17,296	6,987	2,810
Patients			201	44	34

We chose this checkpoint over larger EEG foundation models primarily to avoid data leakage. TUAB, TUEV and TUSZ are all drawn from the TUH EEG Corpus (TUEG), and the most widely used recent foundation models are pretrained on TUEG or its subsets: CBraMod on TUEG itself, whose authors re-pretrain a separate variant without TUEV for this reason (Wang et al., 2025), and LaBraM on a collection that includes TUSZ (Jiang et al., 2024). Their pretraining data may therefore contain our evaluation patients. PREST contains no TUH recordings, so no window of any evaluation patient is seen during pretraining. Among the other public BIOT checkpoints, one is pretrained on data that include TUAB and TUEV and another expects 18 channels (Yang et al., 2023), whereas this one matches our 16-channel bipolar TCP montage (Appendix B.1).

FedGME does not depend on this choice. It only requires a representation h, and its communication cost scales with d_h rather than with the number of encoder parameters (Appendix A.7).

B.3 Client partitions

For every seed, each training patient is assigned to exactly one client, so clients hold disjoint patients, and label skew is then imposed deliberately. On the multiclass corpora, event classes are omitted from selected clients while every class remains at several clients: on TUEV, background (BCKG) is kept at every client and a client lacks on average 2.6 of the six classes at both K; on TUSZ, a client lacks on average 1.2 of the four classes at $K = 5$ and 0.6 at $K = 10$. On TUAB, both labels stay at every client and only their ratio varies across clients. The seed also fixes the model initialisation, so the reported standard deviations reflect partition heterogeneity together with training noise.

Table 5: Composition of the TUH EEG Events Corpus (TUEV) (Obeid & Picone, 2016). Each sample is a 5 s window centred on a 1 s channel-level annotation, using the 16-channel bipolar TCP montage. In our runs, validation is held out at runtime from each client's own training patients (15%), patient-disjoint from training and test.

Label	Class	Description	Train	Val	Test
0	SPSW	Interictal spike or sharp-wave epileptiform transient.	416 (0.61%)	229 (1.46%)	567 (1.93%)
1	GPED	Generalized periodic epileptiform discharges recurring bilaterally.	10,739 (15.73%)	515 (3.29%)	4,677 (15.90%)
2	PLED	Periodic epileptiform discharges localized to one hemisphere.	4,656 (6.82%)	1,528 (9.76%)	1,998 (6.79%)
3	EYEM	Eye-movement artifacts caused by blinks or saccades.	1,011 (1.48%)	59 (0.38%)	329 (1.12%)
4	ARTF	Non-cerebral instrumental, movement, or myogenic artifact.	8,744 (12.81%)	2,309 (14.74%)	2,204 (7.49%)
5	BCKG	Normal background EEG without an annotated event.	42,704 (62.55%)	11,022 (70.37%)	19,646 (66.78%)
Total windows			68,270	15,662	29,421

B.4 BASELINES

Table 6 lists what each method exchanges, what the server does with it, and what each client optimises in the representation phase. All methods start from the same pretrained encoder and are evaluated with the same probe (Appendix B.6); pretrained anchoring is used by FedGME only.

Table 6: Baselines and ablations: exchanged payload, server operation and local objective in the representation phase.

Method	Exchanged per class and round	Server	Local objective
Local	nothing	–	cross-entropy on local labels
FedAvg	all model parameters (instead of class summaries)	uniform parameter average	cross-entropy, end to end
FedPCL	one z-centroid	collects per-client prototypes and their per-class average	prototype-wise contrastive loss (Tan et al., 2022b) against the averaged prototypes and against each client's prototypes, absent classes padded from the average; encoder trained locally
FedPCL-S	one z-centroid	collects per-client prototypes	$\mathcal{L}_{\text{local}}$ (Sec. 3.4) in place of FedPCL's loss, with all $K\|\mathcal{C}\|$ per-client prototypes as weighted positives in $\mathcal{L}_{\text{sup}}$
FedPCL-SM	one z-space Gaussian component	bank, never averaged	$\mathcal{L}_{\text{local}}$ on the own batch $+ \mathcal{L}^{z}_{\text{bank}}$
FedGME	one component in z and one in h	bank, never averaged	$\mathcal{L}_{\text{rep}} + \lambda_{\text{fr}}\mathcal{L}_{\text{freeze}}$, including pretrained anchoring; classifier trained with $\mathcal{L}_{\text{replay}}$

The first step of the ladder, FedPCL to FedPCL-S, changes only the local loss, into which the exchanged prototypes enter as weighted positives. The second, FedPCL-S to FedPCL-SM, replaces the centroid by a z-space Gaussian component and, with it, the prototype term by the bank score. The third, FedPCL-SM to FedGME, adds the h-space bank, pretrained anchoring, absent-subspace anchoring and synthetic replay. Because the last two steps each change more than one element, we attribute the absent-class gains to these mechanisms jointly rather than to any one of them.

B.5 TRAINING DETAILS

All methods share the optimisation setup below; only the loss terms differ (Table 6).

Model. The encoder f_k is BIOT (Yang et al., 2023) with four transformer blocks, eight heads and $d_h = 256$, taking 16-channel input at 200 Hz with token size 200 and hop 100; all four blocks are trainable. The projector p_k maps to $d_z = 64$, and g_k is a linear probe.

Optimiser. AdamW with one parameter group per module and phase (Sec. 3.6). In the representation phase, encoder and projector are trained at learning rate 10^{-4} with weight decay 5×10^{-4}, and g_k, trained alongside through $\mathcal{L}_{\text{replay}}$, with weight decay 5×10^{-5}. In the probe phase, encoder and projector are frozen and g_k is trained with the settings of Appendix B.6.

Schedule. Each round is one local epoch at batch size 64 (256 at evaluation), and optimiser state persists across rounds within a phase. The learning rate ramps up linearly over the first 500 steps of the representation phase, counted from the start of the phase rather than of each round. It then decays on a cosine to half its base value by the last representation round. The ramp protects the pretrained encoder at the start of training: the projector is freshly initialised, so its early gradients are uninformative, and the contrastive objectives are scale-free, so nothing in the loss resists a large first step.

Sampling. Batches are drawn by a class-balanced sampler with exponent 0.75, i.e. class c is sampled in proportion to $n_{kc}^{0.75}$. A per-recording cap prevents a single long recording from filling a batch with crops of itself.

B.6 EVALUATION PROTOCOL

No method fine-tunes its encoder for evaluation. After training, every arm is read out by a linear probe on its frozen encoder features: a single linear layer with one output per class in $\mathcal{C}$, trained for 10 probe rounds with AdamW (learning rate 4×10^{-3}, weight decay 10^{-4}) and balanced class weights, identically for every arm. The probe is evaluated locally on each client using the full common test set of the corpus, and the resulting metrics are averaged across all clients. Differences between rows therefore reflect the learned representations rather than classifier capacity or fine-tuning. Balanced accuracy on locally absent classes (Abs.) is balanced accuracy restricted to the classes a client never trains on, averaged over the clients that hold at least one such class.

C ADDITIONAL RESULTS

C.1 DOWNSTREAM PERFORMANCE

Tables 7–9 report, with standard deviations, the metrics customary for each corpus (Yang et al., 2023). They agree with Table 1 that FedGME is best in every setting but one: on TUEV at $K = 10$, FedAvg leads on Cohen's κ (0.351 vs. 0.267) and F1_w (0.636 vs. 0.596), both of which are dominated by the majority BCKG class (Sec. 4). On TUSZ, FedGME leads on all three metrics at both K, including AUROC (0.653 and 0.649, against 0.597 and 0.632 for FedAvg). On TUAB, its margins on AUPRC and AUROC stay within 0.007 of the next method, consistent with TUAB acting as a control. On TUEV at $K = 5$, FedGME and FedAvg vary more across seeds than the other methods (± 0.035 and ± 0.037 balanced accuracy), plausibly because their scores depend on which classes each partition leaves absent.

C.2 LOCALLY ABSENT CLASSES

Table 10 gives the Abs. columns of Table 1 with standard deviations; Sec. 4 discusses them.

Table 7: TUEV: downstream performance for $K = 5$ and $K = 10$ clients, read out as in Appendix B.6. Mean±std over five seeds; best per K in bold.

Method	BAcc	κ	$F1_w$
$K = 5$			
Local	0.316 ±0.015	0.268 ±0.021	0.582 ±0.013
FedAvg	0.348 ±0.037	0.357 ±0.041	0.642 ±0.023
FedPCL	0.328 ±0.012	0.280 ±0.013	0.604 ±0.010
FedPCL-S	0.336 ±0.009	0.296 ±0.013	0.610 ±0.010
FedPCL-SM	0.327 ±0.010	0.287 ±0.019	0.609 ±0.010
FedGME	**0.408** ±0.035	**0.367** ±0.062	**0.661** ±0.037
$K = 10$			
Local	0.300 ±0.006	0.227 ±0.006	0.555 ±0.005
FedAvg	0.353 ±0.003	**0.351** ±0.011	**0.636** ±0.008
FedPCL	0.319 ±0.019	0.243 ±0.014	0.576 ±0.005
FedPCL-S	0.321 ±0.014	0.248 ±0.004	0.579 ±0.005
FedPCL-SM	0.317 ±0.019	0.257 ±0.005	0.586 ±0.001
FedGME	**0.377** ±0.007	0.267 ±0.011	0.596 ±0.014

Table 8: TUSZ: downstream performance for $K = 5$ and $K = 10$ clients. Protocol as in Table 7.

Method	BAcc	$F1_w$	AUROC
$K = 5$			
Local	0.284 ±0.004	0.407 ±0.007	0.556 ±0.017
FedAvg	0.286 ±0.004	0.410 ±0.008	0.597 ±0.018
FedPCL	0.279 ±0.004	0.402 ±0.009	0.571 ±0.016
FedPCL-S	0.279 ±0.006	0.403 ±0.008	0.569 ±0.021
FedPCL-SM	0.280 ±0.005	0.404 ±0.008	0.567 ±0.023
FedGME	**0.339** ±0.011	**0.500** ±0.013	**0.653** ±0.012
$K = 10$			
Local	0.283 ±0.001	0.522 ±0.005	0.542 ±0.011
FedAvg	0.293 ±0.003	0.539 ±0.002	0.632 ±0.005
FedPCL	0.282 ±0.002	0.527 ±0.001	0.537 ±0.011
FedPCL-S	0.278 ±0.003	0.524 ±0.001	0.534 ±0.021
FedPCL-SM	0.284 ±0.002	0.531 ±0.007	0.556 ±0.001
FedGME	**0.350** ±0.013	**0.554** ±0.009	**0.649** ±0.019

C.3 Representation geometry

Table 11 extends panel (b) of Table 1 to all corpora. Effective rank (Roy & Vetterli, 2007; Garrido et al., 2023) and uniformity (Wang & Isola, 2020) are computed on the uncentred, ℓ_2-normalised encoder embeddings $\hat{h} \in \mathbb{R}^{256}$.

Table 9: TUAB: downstream performance for $K = 5$ and $K = 10$ clients. Protocol as in Table 7.

Method	BAcc	$F1_w$	AUPRC	AUROC
$K = 5$				
Local	0.721 ±0.005	0.723 ±0.004	0.782 ±0.003	0.794 ±0.002
FedAvg	0.742 ±0.008	0.745 ±0.009	0.800 ±0.011	0.808 ±0.010
FedPCL	0.742 ±0.001	0.743 ±0.002	0.796 ±0.001	0.811 ±0.001
FedPCL-S	0.748 ±0.005	0.750 ±0.003	0.804 ±0.003	0.816 ±0.003
FedPCL-SM	0.747 ±0.003	0.749 ±0.004	0.801 ±0.004	0.814 ±0.002
FedGME	**0.749** ±0.002	**0.752** ±0.002	**0.811** ±0.002	**0.822** ±0.003
$K = 10$				
Local	0.716 ±0.007	0.718 ±0.008	0.775 ±0.001	0.787 ±0.002
FedAvg	0.737 ±0.010	0.739 ±0.010	0.788 ±0.015	0.801 ±0.012
FedPCL	0.729 ±0.003	0.731 ±0.003	0.791 ±0.002	0.803 ±0.005
FedPCL-S	0.732 ±0.003	0.734 ±0.003	0.787 ±0.003	0.800 ±0.003
FedPCL-SM	0.730 ±0.003	0.731 ±0.004	0.790 ±0.004	0.801 ±0.003
FedGME	**0.740** ±0.006	**0.742** ±0.006	**0.798** ±0.004	**0.809** ±0.004

Table 10: Balanced accuracy on locally absent classes: for each client, BAcc restricted to the classes that client never trains on, averaged over the clients holding at least one such class and over five seeds (± one standard deviation). TUEV has on average 2.6 absent classes per client; TUSZ has 1.2 at $K = 5$ and 0.6 at $K = 10$; TUAB has none.

Method	TUEV	TUSZ
$K = 5$		
Local	0.000 ±0.001	0.012 ±0.002
FedAvg	0.103 ±0.012	0.054 ±0.003
FedPCL	0.000 ±0.000	0.010 ±0.001
FedPCL-S	0.000 ±0.000	0.011 ±0.003
FedPCL-SM	0.000 ±0.000	0.011 ±0.006
FedGME	**0.164** ±0.018	**0.150** ±0.033
$K = 10$		
Local	0.000 ±0.000	0.009 ±0.010
FedAvg	0.104 ±0.007	0.000 ±0.000
FedPCL	0.000 ±0.000	0.011 ±0.008
FedPCL-S	0.000 ±0.000	0.011 ±0.011
FedPCL-SM	0.000 ±0.000	0.011 ±0.009
FedGME	**0.183** ±0.013	**0.295** ±0.042

Table 11: Representation geometry of the normalised encoder embeddings for the runs of Table 1. eRank: effective rank (at most 256); Unif.: uniformity (lower is more uniform). Mean±std over five seeds.

	TUEV		TUSZ		TUAB	
Method	eRank	Unif. ↓	eRank	Unif. ↓	eRank	Unif. ↓
$K = 5$						
Local	78.1 ±2.6	−1.56 ±0.04	82.8 ±0.9	−1.64 ±0.05	97.6 ±1.8	−1.90 ±0.06
FedAvg	100.8 ±3.6	−2.00 ±0.07	106.5 ±1.3	−2.79 ±0.05	127.0 ±1.2	−2.91 ±0.00
FedPCL	50.8 ±4.5	−1.22 ±0.12	88.8 ±1.7	−2.32 ±0.08	41.3 ±8.9	−1.01 ±0.06
FedPCL-S	101.2 ±1.0	−3.10 ±0.06	88.6 ±1.1	−2.55 ±0.07	142.5 ±2.0	−3.21 ±0.08
FedPCL-SM	76.2 ±4.5	−2.39 ±0.17	89.1 ±1.1	−2.49 ±0.03	139.8 ±3.3	−3.30 ±0.05
FedGME	74.8 ±1.1	−1.77 ±0.01	98.9 ±0.4	−2.87 ±0.02	124.2 ±1.8	−2.94 ±0.05
$K = 10$						
Local	73.5 ±0.9	−1.74 ±0.02	88.1 ±1.3	−1.95 ±0.03	96.4 ±0.4	−1.96 ±0.01
FedAvg	89.4 ±2.2	−2.25 ±0.10	99.9 ±0.5	−2.60 ±0.00	127.0 ±1.0	−3.15 ±0.01
FedPCL	51.8 ±2.3	−1.42 ±0.10	81.9 ±1.4	−2.21 ±0.07	32.1 ±1.8	−0.93 ±0.04
FedPCL-S	87.5 ±1.0	−2.94 ±0.03	82.3 ±1.3	−2.39 ±0.06	134.4 ±1.8	−3.19 ±0.04
FedPCL-SM	76.4 ±0.6	−2.59 ±0.01	84.3 ±0.1	−2.48 ±0.01	130.0 ±1.3	−3.27 ±0.04
FedGME	63.9 ±0.3	−1.71 ±0.04	97.8 ±0.3	−2.89 ±0.01	119.0 ±0.6	−2.76 ±0.00