Quantifying Domain Gaps for RF Spectrogram Classification: A Leave-One-Domain-Out and k-NN Agreement Study

Gabriel Urbaitis and Andrew Cochrane
COSMIAC Research and Development Center
University of New Mexico, Albuquerque, NM, USA
Gabriel.Urbaitis@cosmiac.org, Andrew.Cochrane@cosmiac.org

***Abstract*—Deployed RF classifiers often face *domain shift*: data at test time differ from data used to design the model, producing generalization failures [1]. This paper presents a compact, training-free protocol that measures such gaps directly in feature space for a six-class spectrogram task with three sources: a bench-generated *Lab Source* and two independent field collections (*Field Source A*, *Field Source B*). Using ImageNet-pretrained ResNet-18 features with PCA(50), we quantify cross-domain separability via (i) *leave-one-domain-out* (LODO) 1-nearest-neighbor (1-NN) accuracy and (ii) *cross-domain 10-NN label agreement*. We complement these with unsupervised class×domain clustering (PCA+KMeans), a silhouette summary, and per-class Adjusted Rand Index (ARI) between clusters and domain labels.**

Within-domain 1-NN accuracies are high (Lab $99.27\,\%$, Field A $98.61\,\%$, Field B $96.48\,\%$) yet drop sharply in LODO Lab→A ($43.62\,\%$) and Lab→B ($34.33\,\%$). Cross-domain 10-NN agreement mirrors this pattern. At $k = 22$ (matching the number of observed class×domain pairs), KMeans on PCA(50) yields a silhouette ≈ 0.425; per-class ARI shows strong domain-tie for BOC/CW/BPSK (up to 0.93) and only *moderate* domain-tie for AWGN/Chirp/No Jammer (0.31–0.36). These findings align with domain generalization principles [2] and motivate multi-source training. The analyses are simple to reproduce, require no model training, and provide actionable guidance on *what to collect next*.

***Index Terms*—RF spectrograms, domain shift, LODO, nearest neighbors, clustering, domain generalization.**

I. Introduction

Differences in hardware front-ends, propagation conditions, interference backgrounds, sampling regimes, and preprocessing pipelines commonly induce *dataset shift* between development and deployment [1]. In RF perception tasks (e.g., jammer/modulation classification from spectrograms), this shift can substantially reduce accuracy even when the label set is unchanged.

This paper focuses on **measuring** domain gaps in a way that is light-weight, transparent, and *training-free*. Rather than fit large models and disentangle multiple confounds, we analyze the geometry of generic visual features (ResNet-18 penultimate activations) as a proxy for how a wide class of CNN-like encoders will view the data manifold. Our three sources include a **Lab Source** (bench/controlled) and two independent **Field** sources (**A**, **B**), across six classes.

Contributions

- A concise *LODO 1-NN* protocol that quantifies cross-domain separability without training a classifier [3].
- A complementary *10-NN cross-domain label-agreement* metric that measures neighborhood label consistency across domains.
- Unsupervised class×domain analysis (PCA+KMeans) summarizing structure via silhouette and per-class ARI.
- Practical guidance: when LODO gaps are large, multi-source training or domain-robust objectives [2] are justified; our per-class diagnostics indicate where to target effort.

II. Related Work (brief)

Dataset Shift. Quionero-Candela et al. [1] formalize types of shift (covariate, prior, concept) and discuss evaluation pitfalls. Our analysis is an *evaluation tool*: it does not assume which type of shift is present; it empirically measures separability.

Domain Generalization. Gulrajani and Lopez-Paz [2] analyze algorithmic and evaluation choices for domain generalization (DG). Our results complement DG by diagnosing whether multi-source training is warranted and which classes need invariance the most.

Nearest-Neighbor Reasoning. Cover and Hart [3] showed the asymptotic optimality of 1-NN under mild conditions. Here 1-NN is not a deployed classifier; it is a transparent *probe* of neighborhood consistency across domains.

III. Data and Feature Space

We consider six classes typical in RF spectrogram analysis.[1] The three domains are **Lab Source** and **Field Source A/B**. Unless noted, we restrict to *clean* (no additive synthetic noise) spectrograms.

[1]Exact class names are standard (e.g., BOC, BPSK, CW, Chirp, AWGN, No Jammer); our labels follow that convention.

A. Feature extraction

Each image is resized to 224×224 and embedded with ResNet-18 (ImageNet pretrained) to a 512-D feature. We apply PCA to 50 dimensions (whitened) for distance computations and clustering; PCA(2) is used only for visualization (Fig. 1). Cosine distance is used for k-NN; Euclidean gave qualitatively similar results in spot checks.

B. Sampling and scale

To keep compute modest and avoid heavy class imbalance, we cap to at most 1,500 images per class across domains for a total of $N = 9{,}000$ samples in the nearest-neighbor and clustering analyses.

IV. Protocols and Metrics

Let $D \in \{\text{Lab}, \text{A}, \text{B}\}$ index domains; $X_D = \{(x_i^{(D)}, y_i^{(D)})\}$ with x in PCA(50) and label $y \in \{1, \ldots, 6\}$. Let $d(\cdot, \cdot)$ be cosine distance.

A. LODO 1-NN accuracy

Given a *query* domain Q and a *gallery* domain G,

$$\mathrm{NN}_G(q) = \arg\min_{g \in X_G} d(q, g), \tag{1}$$

$$\hat{y}(q) = y\big(\mathrm{NN}_G(q)\big), \tag{2}$$

$$\mathrm{Acc}_{1\text{-NN}}(Q{\to}G) = \Pr\big[\hat{y}(q) = y(q)\big]_{q \sim X_Q}. \tag{3}$$

Within-domain uses $Q{=}G$ but excludes trivial self-matches by requiring $\mathrm{NN}_G(q)$ to be a *different* sample.

B. Cross-domain k-NN agreement (here k=10)

For each $q \in X_Q$, retrieve $\mathcal{N}_{10,G}(q)$: the 10 nearest neighbors of q in X_G. Define

$$\mathrm{Agree}_{10}(Q{\to}G) = E_{q \sim X_Q}\left[\frac{1}{10} \sum_{g \in \mathcal{N}_{10,G}(q)} \mathbf{1}\{y(g) = y(q)\}\right]. \tag{4}$$

This is not a k-NN *classifier*; it reports the local label homophily across domains.

C. Unsupervised class×domain analysis

We run KMeans on PCA(50). We examine:

- **Silhouette** to summarize cluster separation.
- **Per-class ARI** between domain labels and cluster ids (restricting to samples of that class): high ARI means the class partitions by domain.
- **Fragmentation** counts: number of clusters where a class exceeds 10% of cluster mass (*how scattered* a class is).

V. Results

A. Nearest-neighbor domain gaps

Table I shows very strong within-domain agreement but large LODO drops from Lab to both field domains. This indicates that neighborhoods reorganize across domains, even for generic visual features.

TABLE I
LODO 1-NN accuracy and within-domain baselines (PCA(50), cosine).

Metric	Accuracy (%)
Within-domain (Lab)	99.27
Within-domain (Field A)	98.61
Within-domain (Field B)	96.48
LODO Lab→Field A	43.62
LODO Lab→Field B	34.33

TABLE II
Cross-domain 10-NN label agreement (mean fraction of neighbors sharing the query label).

Direction	Agreement (%)
Lab→Field A	40.60
Lab→Field B	36.56
Field A→Lab	47.53
Field B→Lab	35.39

B. Cross-domain 10-NN label agreement

Agreement trends echo LODO (Table II). Note the asymmetry (e.g., Field A→Lab vs. Lab→A), which is typical when domains differ in diversity and noise floors.

C. Class-by-class domain sensitivity

At $k = 22$ (matching the number of observed class×domain pairs), KMeans on PCA(50) yields a silhouette ≈ 0.425. Per-class ARI (cluster vs. domain) reveals uneven domain pull (Table III): BOC is extremely domain-tied (0.93), CW/BPSK are strongly domain-tied $(0.72/0.61)$, while Chirp/AWGN/No Jammer are only moderately domain-tied $(0.34/0.31/0.36)$. Consistent with this, *fragmentation* at $k{=}22$ shows which classes spread across many clusters.

Fragmentation at $k = 22$.: Number of clusters where each class exceeds 10% of cluster mass: *Chirp 9, AWGN 8, BOC 5, BPSK 5, CW 4, No Jammer 3*. Higher counts indicate a class spreads over more clusters (greater heterogeneity).

D. Sensitivity: varying k in KMeans

We sweep $k \in \{6, 8, 12, 16, 20, 22, 24\}$ to examine the silhouette/ARI trade-off (Table IV). Two patterns are useful operationally: (i) $k{=}20$ maximizes silhouette (0.464) and gives the best *class*-alignment (ARI_class $= 0.324$), suggesting more compact, class-oriented clusters; (ii) $k{=}22$ maximizes alignment with *class×domain* structure (ARI_clsDom $= 0.492$), making domain fragmentation most explicit. We therefore report $k{=}22$ in Fig. 1 when the goal is to visualize domain-specific splitting, and use the sweep to show robustness.

E. Visualization

Fig. 1 shows PCA(2) colored by KMeans cluster, annotated with majority (class — domain). Clear domain-specific splits are visible for several classes.

TABLE III
PER-CLASS ARI BETWEEN CLUSTER ASSIGNMENTS AND DOMAIN LABELS AT $k = 22$ (HIGHER MEANS MORE DOMAIN-TIED).

Class	ARI
BOC	0.93
CW	0.72
BPSK	0.61
Chirp	0.34
AWGN	0.31
No Jammer	0.36

TABLE IV
KMEANS SENSITIVITY (PCA(50) FEATURES). SILHOUETTE, ARI_CLASS (CLUSTERS VS. GROUND-TRUTH CLASSES), AND ARI_CLSDOM (CLUSTERS VS. CLASS×DOMAIN PAIRS) ACROSS k.

k	Silhouette	ARI_class	ARI_clsDom
6	0.408	0.213	0.158
8	0.392	0.282	0.213
12	0.456	0.264	0.234
16	0.441	0.321	0.321
20	0.464	0.324	0.340
22	0.425	0.290	0.492
24	0.438	0.261	0.393

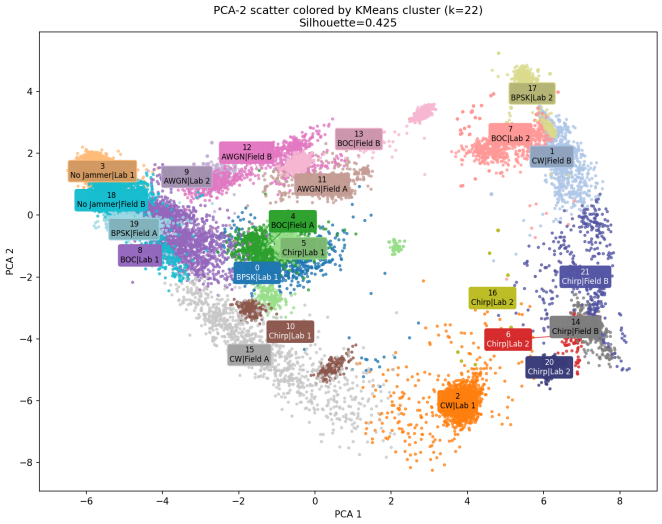

Fig. 1. PCA(2) + KMeans. Each point is a spectrogram feature; labels indicate cluster id and majority (class — domain). The Lab Source pools two bench-generated collections, Lab 1 and Lab 2, for all metrics; labels may show Lab 1/2 to indicate which subset dominates a cluster.

VI. ANALYSIS AND OPERATIONAL GUIDANCE

A. Why multi-source training?

The LODO drops in Table I and low cross-domain agreement in Table II show that neighborhoods reorganize across domains. Training on a single domain risks learning decision boundaries that align with domain-specific artifacts rather than class-defining structure. Mixing *heterogeneous* sources during training widens the support and typically improves robustness [2].

B. Which classes to prioritize?

Per-class ARI (Table III) identifies which classes are most domain-tied. For domain-tied classes (BOC/CW/BPSK), prioritize:

1) targeted cross-domain augmentation (e.g., synthetic channel/FR interference),
2) domain-adversarial or invariant representation objectives during training,
3) feature-space normalization consistent across domains.

For relatively domain-light classes (Chirp/AWGN/No Jammer), simpler regularization often suffices.

C. What data to collect next?

If the intended deployment resembles Field A or Field B, collect additional examples of the domain-tied classes in those environments. The proposed metrics can be rerun incrementally to confirm shrinking gaps as data accrue.

VII. THREATS TO VALIDITY

Feature proxy. We use ResNet-18 (vision) as a generic feature proxy; RF-specific encoders (or self-supervised pretraining on unlabeled RF images) may reduce measured gaps, but large separations typically persist unless explicitly addressed.

Sampling. We cap per-class counts to bound runtime. Different caps can shift numeric values slightly; qualitative patterns (LODO vs. within-domain, class ARI ordering) were stable in our spot checks.

Distance/whitening choices. Cosine with PCA(50) is our default; Euclidean and small changes to PCA dimension did not alter conclusions.

Clustering k. Choosing k to match class×domain pairs emphasizes domain fragmentation; however, the sweep in Table IV shows consistent behavior and clarifies trade-offs.

VIII. CONCLUSION

Simple, training-free probes in feature space reveal substantial domain gaps for RF spectrogram classification. Within-domain nearest neighbors are almost always of the same class, yet cross-domain neighborhoods often disagree. Unsupervised clustering pinpoints which classes are most domain-tied. These diagnostics justify multi-source training and help target effort where it matters.

REPRODUCIBILITY NOTES

ResNet-18 features (ImageNet), PCA(50) for metrics, PCA(2) for plots, cosine distance for k-NN; k=10 for agreement; KMeans with k equal to the count of observed class — domain pairs. Figures can be produced from the generated CSVs and images.

REFERENCES

[1] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press, 2009.

[2] I. Gulrajani and D. Lopez-Paz, "In Search of Lost Domain Generalization," in *ICLR*, 2021.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Information Theory*, 1967.