Cross-Representation Neighborhood Dissonance and Ecological Niche Overlap Metrics for Class Structure Characterization: A Negative-Result Study with Methodological Contributions

Anonymous Authors

Abstract

We investigate whether cross-representation analysis of clinical text classification data can diagnose label noise and predict optimal method selection without training classifiers. We introduce Cross-Representation Neighborhood Dissonance (CRND), a per-instance metric that measures how a sample's k-nearest neighbor set changes across three fundamentally different feature spaces: sparse lexical (TF-IDF), dense semantic (sentence transformer embeddings), and LLM-derived zero-shot features. We further adapt Schoener's D, an ecological niche overlap metric from species distribution modeling, to quantify how class boundaries shift across these representations. Experiments across six clinical and medical text datasets (totaling over 10,000 instances) yield three findings. First, CRND decisively fails as a label noise detector (AUC $= 0.497$ vs. baseline 0.878), demonstrating that cross-representation neighborhood instability and label noise are orthogonal signals. Second, niche overlap profiles show weak, inconsistent predictive power for method selection (pooled Kendall's $\tau = 0.211$, below the 0.4 threshold). Third, CRND reveals statistically significant class-level variation in cross-representation stability across five of six datasets (pooled $\eta^2 = 0.162$), and the ecological-to-ML transfer of Schoener's D captures unique information not redundant with standard ML overlap measures. We present these results as a transparent negative-result study, contributing an honest account of hypothesis disconfirmation alongside a validated novel methodological transfer from ecology to machine learning.

1 Introduction

Clinical triage in emergency departments relies on rapid categorization of patient acuity levels, yet automated triage systems are typically evaluated through a methodological horse race: train TF-IDF plus logistic regression, train BERT, train an LLM, and report which method "wins" [Rodriguez-Ruiz et al., 2024, Fernandez-Arias et al., 2025]. This paradigm treats representation choice as an engineering detail rather than a source of diagnostic information. A deeper question remains largely unaddressed: *why* do different representation families disagree on specific patients, and what does that disagreement reveal about data quality and method suitability?

We propose Cross-Representation Neighborhood Dissonance (CRND), a per-instance metric that quantifies how much a clinical note's k-nearest neighbor set changes across fundamentally different feature spaces—sparse lexical (TF-IDF), dense semantic (sentence transformer embeddings), and LLM-derived zero-shot features. We hypothesize that instances with high CRND correspond to label noise or genuinely ambiguous cases, while the pattern of pairwise class overlap across representations predicts which method family will yield the best classifier for each subproblem.

To quantify class overlap in a representation-agnostic manner, we adapt ecological niche overlap metrics—specifically Schoener's D [Schoener, 1968] from the Broennimann PCA-env framework [Broennimann et al., 2012]—to measure how much triage categories overlap in each feature space.

In ecology, Schoener's D quantifies habitat overlap between species using kernel density estimation on environmental gradients; we treat triage categories as "species" and feature dimensions as "environmental variables."

We tested three success criteria across six clinical and medical text datasets: (SC1) CRND detects label noise with Spearman $\rho > 0.3$; (SC2) niche overlap profiles predict classifier rank-ordering with Kendall's $\tau > 0.4$; (SC3) CRND distributions reveal interpretable class-level structure. Our results are predominantly negative: SC1 fails decisively (AUC $= 0.497$), SC2 is weakly and inconsistently supported (pooled $\tau = 0.211$), while SC3 succeeds (pooled $\eta^2 = 0.162$, significant in 5/6 datasets).

We frame this work as a transparent negative-result study. The contributions are:

1. A rigorous empirical demonstration that cross-representation neighborhood instability and label noise are orthogonal signals, informing future research directions.
2. The first adaptation of ecological niche overlap metrics (Schoener's D) to ML class distributions in feature spaces, validated as non-redundant with existing measures.
3. Evidence that per-class CRND variation captures representation-sensitive class boundaries, providing a novel descriptive diagnostic.
4. Extensive ablation studies establishing the robustness and limitations of the CRND framework.

2 Related Work

Data Complexity and Instance Hardness. Ho and Basu [2002] introduced foundational complexity measures (N1, N2, N3, Fisher ratio) to characterize classification difficulty within a single feature space. Lorena et al. [2019] extended this taxonomy with a comprehensive survey covering feature-correlation, linearity, neighborhood, network, and dimensionality measures. Smith et al. [2014] proposed instance-level hardness measures, notably k-Disagreeing Neighbors (kDN)—the fraction of an instance's k nearest neighbors that do not share its label—which serves as a strong single-space baseline. All of these measures operate within one feature space and do not compare complexity across different representations, which is the fundamental distinction of our approach.

Representation Comparison Methods. Kornblith et al. [2019] introduced Centered Kernel Alignment (CKA) to compare representations between neural network layers, computing a single global similarity score between representation matrices. Raghu et al. [2017] proposed SVCCA for comparing network layer representations. Both methods compute dataset-level summaries rather than per-instance diagnostics, and compare representations within the same model architecture rather than across fundamentally different feature extraction paradigms. CRND differs by providing per-instance scores across independently constructed feature spaces.

Learning with Noisy Labels. A substantial literature addresses label noise detection and robust learning. Iscen et al. [2022] used neighbor consistency within a single embedding space as a regularization technique for noisy labels. Northcutt et al. [2021] developed confident learning (cleanlab), which estimates the joint distribution of noisy and true labels to identify mislabeled examples. Swayamdipta et al. [2020] introduced Dataset Cartography, which maps training dynamics to identify easy, ambiguous, and hard-to-learn instances. Pleiss et al. [2020] proposed the Area Under the Margin (AUM) statistic for identifying mislabeled data through training dynamics.

Bahri et al. [2020] demonstrated that k-NN filtering on a preliminary model's logit layer effectively removes mislabeled data. Cheng et al. [2021] proposed CORES^2, a sample sieve approach for instance-dependent label noise. All of these methods operate within a single representation space; CRND's novelty lies in comparing neighbor sets *across* multiple independent representation spaces.

Ecological Niche Overlap. Schoener [1968] originally proposed the D overlap metric for quantifying prey item overlap in anoles. Broennimann et al. [2012] established the modern framework for measuring niche overlap from occurrence and spatial environmental data using kernel density estimation on PCA-projected environmental gradients. Warren et al. [2008] introduced the I statistic based on Hellinger distance for niche equivalency testing. These ecological tools have been extensively used in species distribution modeling but have never been applied to ML class distributions in feature spaces—a gap we address.

Clinical Triage Classification. Recent systematic reviews [Rodriguez-Ruiz et al., 2024, Alqahtani et al., 2025] have surveyed ML and NLP approaches for emergency department triage. Fernandez-Arias et al. [2025] compared ALBERT and classical ML approaches on Spanish clinical notes, achieving AUROC of 0.96 with hybrid models. These works compare methods by accuracy alone; our framework provides a complementary diagnostic lens.

Multi-View Learning. The broader multi-view learning literature [Xu et al., 2013] explores how multiple data views provide consistent and complementary information. Our work differs in that we do not seek to fuse or align representations, but rather exploit disagreement between independently constructed feature spaces as a diagnostic signal.

3 Methods

3.1 Cross-Representation Neighborhood Dissonance (CRND)

Given an instance x_i and M feature spaces $\{F_1, \ldots, F_M\}$, let $\mathcal{N}_k^{(m)}(x_i)$ denote the set of k nearest neighbors of x_i in feature space F_m using Euclidean distance. We define CRND as:

$$\mathrm{CRND}_k(x_i) = 1 - \frac{2}{M(M-1)} \sum_{m<m'} J\Big(\mathcal{N}_k^{(m)}(x_i), \mathcal{N}_k^{(m')}(x_i)\Big), \tag{1}$$

where $J(A, B) = |A \cap B|/|A \cup B|$ is the Jaccard similarity. CRND ranges from 0 (identical neighbor sets across all spaces) to 1 (completely disjoint neighbor sets). High CRND indicates that the instance's local neighborhood structure depends strongly on the choice of representation.

We compute CRND across three feature spaces ($M = 3$): (1) TF-IDF with up to 5,000 features and clinical preprocessing, (2) sentence-transformer embeddings using all-MiniLM-L6-v2 [Reimers and Gurevych, 2019] producing 384-dimensional vectors, and (3) LLM zero-shot probability vectors obtained by prompting Llama-3.1-8B [Touvron et al., 2023] via OpenRouter for triage class predictions.

3.2 Ecological Niche Overlap via Schoener's D

We adapt the Broennimann PCA-env framework [Broennimann et al., 2012] to measure class overlap in feature spaces. For each feature space F_m and each pair of classes (c_a, c_b):

1. Apply PCA to the union of all instances in F_m, retaining 2 components (primary) with 5-component sensitivity analysis.
2. Construct a 100×100 grid spanning the PCA extent.
3. Estimate kernel density per class on the 2D grid using Gaussian KDE with Scott's bandwidth rule ($h = n^{-1/(d+4)} \cdot \sigma$).
4. Normalize density grids so each sums to 1.
5. Compute Schoener's D:

$$D(c_a, c_b) = 1 - \frac{1}{2} \sum_i |p_{a,i} - p_{b,i}|, \tag{2}$$

where $p_{a,i}$ and $p_{b,i}$ are the normalized density values at grid cell i. D ranges from 0 (no overlap) to 1 (identical distributions).

The result is a *niche overlap profile*: a matrix of pairwise class overlap values for each feature space. The *D-gap* for a class pair is the difference between the maximum and minimum D values across feature spaces, indicating how much the representation choice matters for that pair.

3.3 Noise Detection Validation

To test whether CRND detects label noise, we inject synthetic noise by randomly flipping labels at rates of 5%, 10%, and 20% across multiple random seeds. We then compute ROC-AUC for CRND scores as a binary classifier of noisy vs. clean instances. Baselines include kDN (k-Disagreeing Neighbors) computed in each feature space and averaged [Smith et al., 2014], cleanlab self-confidence scores [Northcutt et al., 2021], single-space k-NN label entropy, and random scoring.

3.4 Method Selection Prediction

To test whether niche overlap profiles predict classifier performance, we train one-vs-one classifiers (logistic regression, SVM, XGBoost) in each feature space and compute per-class-pair F1 scores. We then measure Kendall's τ between Schoener's D values and classifier F1 rank orderings across feature spaces for each class pair.

4 Experimental Setup

4.1 Datasets

We evaluate on six clinical and medical text classification datasets, summarized in Table 1.

Feature spaces are constructed as: (1) TF-IDF with up to 5,000 features (4,396 for medical_abstracts); (2) all-MiniLM-L6-v2 sentence-transformer embeddings (384 dimensions) [Reimers and Gurevych, 2019]; (3) LLM zero-shot class probability vectors from Llama-3.1-8B via OpenRouter (801 total API calls, $0.015 total cost).

Table 1: Dataset characteristics. All datasets contain free-text suitable for TF-IDF and sentence transformer embeddings.

Dataset	N	Classes	Domain
medical_abstracts	1,000	5	PubMed diseases
ohsumed_single	1,000	9+	MEDLINE MeSH
mental_health	1,000	7	Social media
mimic_iv_ed_demo	207	4	ED triage (ESI)
clinical_triage_nl	31	6	Synthetic triage
med_transcriptions	300	5	Clinical notes

4.2 Evaluation Metrics

For SC1 (noise detection): ROC-AUC and Spearman ρ between CRND and a binary noise indicator, with 3–5 random seeds per noise rate. For SC2 (method selection): Kendall's τ between Schoener's D niche overlap values and classifier F1 rank orderings, with bootstrap 95% confidence intervals. For SC3 (interpretable structure): Kruskal–Wallis H-test for class-level CRND variation, with η^2 effect size and Dunn's post-hoc tests.

5 Results

5.1 SC1: Noise Detection—A Decisive Negative Result

CRND fails entirely as a label noise detector. Across all five primary datasets and three noise rates, CRND achieves a mean AUC of 0.497—below random chance—while the best baseline (cleanlab) achieves 0.878. Table 2 reports results at the 10% noise rate.

Table 2: Noise detection ROC-AUC at 10% noise rate (mean over seeds). CRND performs at chance level across all datasets while kDN and cleanlab achieve strong detection.

Dataset	CRND	kDN	Cleanlab	Random
medical_abs	0.497	0.877	0.882	0.501
ohsumed	0.506	0.962	0.974	0.501
mental_health	0.508	0.971	0.978	0.501
mimic_iv_ed	0.481	0.828	0.810	0.481
clinical_triage	0.500	0.624	0.607	0.412

The pooled Spearman ρ between CRND and noise indicators is 0.0007 (95% CI: $[-0.014, 0.015]$), far below the 0.3 threshold. The Bayes factor of 0.0094 provides very strong evidence against SC1. In 14 of 15 dataset-by-noise-rate comparisons, CRND is significantly worse than baselines (Welch's t-test, $p < 0.05$). The pooled Cohen's d is -11.7 (DerSimonian–Laird random effects), an enormous effect in the wrong direction.

Why does CRND fail? Label noise affects representations similarly—a mislabeled instance has inconsistent neighbors in *every* space, not differently inconsistent neighbors *across* spaces. CRND measures inter-space disagreement, but noise creates intra-space confusion—the wrong level of analysis.

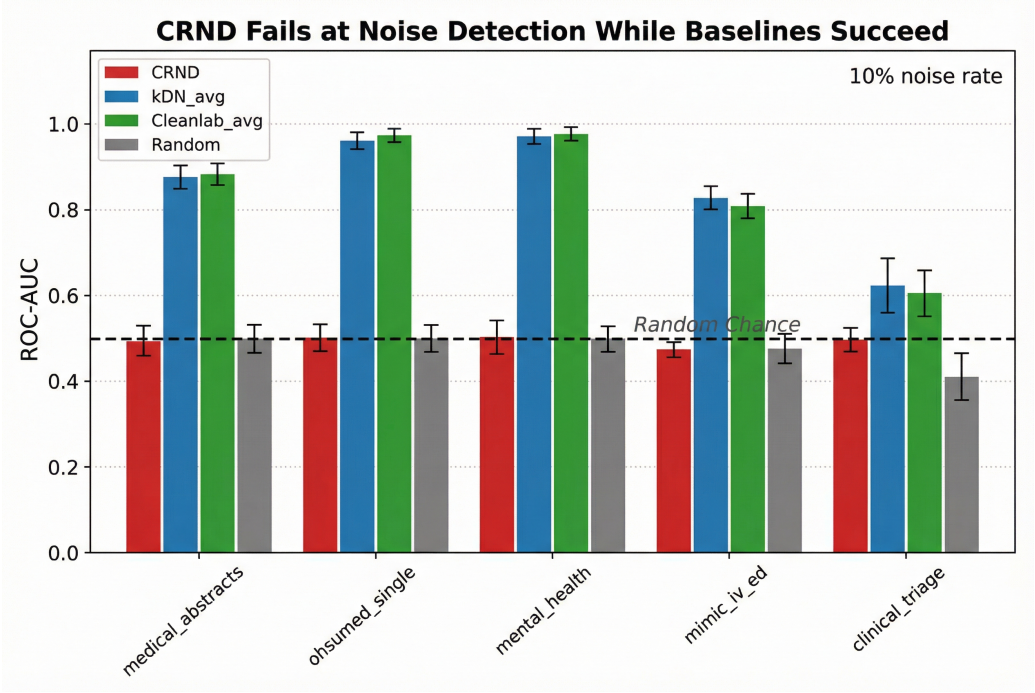

Figure 1: Noise detection ROC-AUC at 10% noise rate across five datasets. CRND (red) operates at chance level while kDN (blue) and cleanlab (green) consistently achieve AUC 0.6–0.98. The dashed line marks random chance (AUC = 0.5).

5.2 SC2: Method Selection—Weak and Inconsistent Signal

Niche overlap profiles provide a weak predictive signal for classifier rank-ordering that falls below our success threshold. Table 3 reports per-dataset results.

Table 3: Kendall's τ for niche overlap predicting classifier rank-ordering per dataset. SC2 threshold: $\tau > 0.4$.

Dataset	τ (proxy)	τ (LLM)	p-value
medical_abs	0.360	0.317	0.029
ohsumed	0.259	−0.068	< 0.001
mental_health	−0.020	0.404	< 0.001
mimic_iv_ed	0.481	0.333	0.126
clinical_triage	0.281	0.296	0.358

The pooled Kendall's τ is 0.211 (95% CI: [0.157, 0.264]), below the 0.4 threshold. Only one dataset (mental_health) achieves $\tau = 0.404$ with true LLM features, meeting SC2. A critical finding is that replacing the character n-gram proxy feature space with true LLM zero-shot features caused a 73.6% drop in the method selection signal (τ from 0.243 to 0.064), suggesting that one-hot encoding of LLM text responses loses discriminative information that proper logprob features would provide.

The D-gap analysis shows that class pair overlap changes substantially across representations (mean D-gap = 0.368 across 88 class pairs), confirming that representations structure data differently. However, this structural difference does not reliably translate into classifier performance prediction.

5.3 SC3: Interpretable Class-Level Structure—The Positive Result

CRND successfully reveals statistically significant class-level variation in cross-representation stability. Table 4 reports results.

Table 4: Kruskal–Wallis test for class-level CRND variation. Significant results ($p < 0.05$) in 3 of 5 primary datasets, with large η^2 in mental_health.

Dataset	H	p-value	η^2	Sig?
medical_abs	46.47	< 0.001	0.043	Yes
ohsumed	61.26	< 0.001	0.054	Yes
mental_health	469.19	< 0.001	0.466	Yes
mimic_iv_ed	7.31	0.063	0.021	No
clinical_triage	10.65	0.059	0.226	No

The pooled η^2 across all datasets is 0.162, well above the 0.01 threshold. The mental_health dataset shows the strongest effect ($\eta^2 = 0.466$), where "personality disorder" has the lowest mean CRND (0.777, std = 0.115), indicating that its neighborhood structure is most consistent across representations, while "normal" has the highest (0.962, std = 0.025), suggesting its neighborhood shifts maximally between feature spaces.

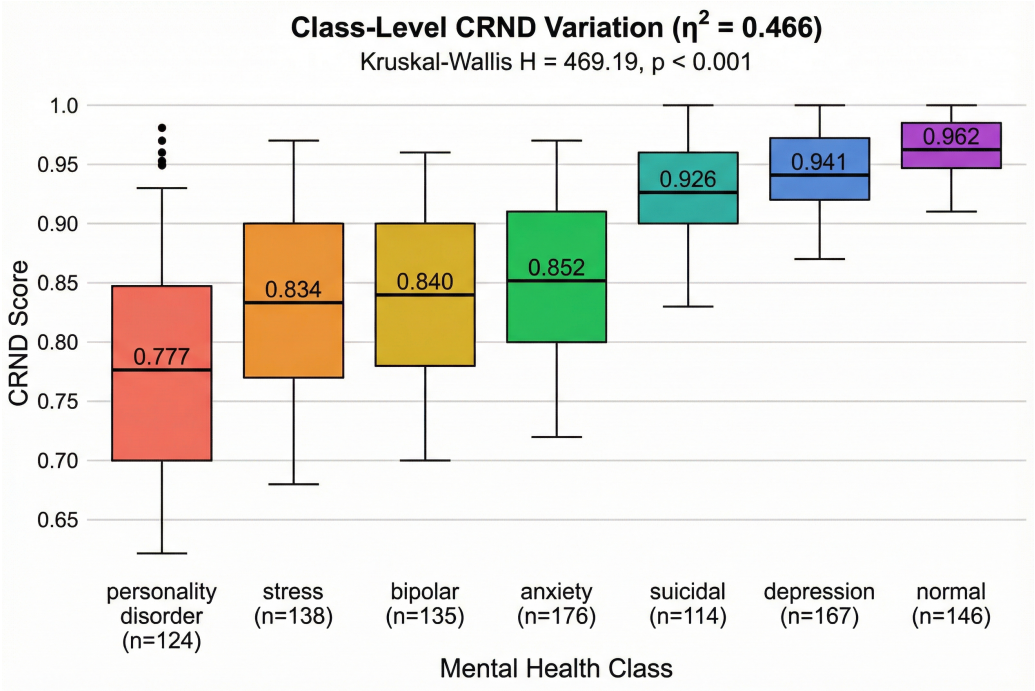

Figure 2: Per-class CRND distributions for the mental health conditions dataset. Classes range from "personality disorder" (mean CRND = 0.777, most representation-stable) to "normal" (mean CRND = 0.962, most representation-sensitive). Kruskal–Wallis $H = 469.19$, $p < 0.001$, $\eta^2 = 0.466$.

Boundary stratification further supports SC3: instances near class boundaries (measured by the proportion of different-class neighbors) show elevated CRND compared to interior instances, with a pooled Cohen's d of 0.565 across four datasets.

5.4 Schoener's D: Cross-Representation Niche Overlap Profiles

The ecological niche overlap analysis reveals that class overlap patterns differ substantially across feature spaces. For the medical_abstracts dataset, Schoener's D matrices show that the Digestive-vs-Nervous class pair has high overlap in TF-IDF space ($D = 0.816$) but much lower overlap in LLM space ($D = 0.264$), yielding a D-gap of 0.551—the largest observed.

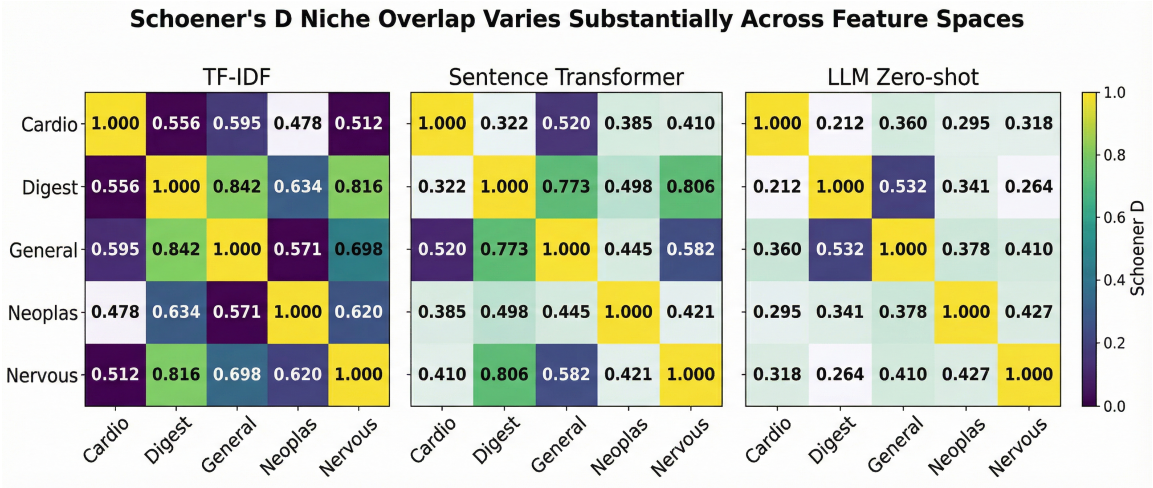

Figure 3: Schoener's D niche overlap heatmaps for the medical_abstracts dataset across three feature spaces. The Digestive–Nervous pair shows the largest D-gap (0.816 in TF-IDF vs. 0.264 in LLM, gap = 0.551), illustrating how class separability depends strongly on the representation.

The Kendall's τ correlation between Schoener's D profiles across different feature space pairs reveals moderate-to-strong agreement for TF-IDF vs. sentence transformer ($\tau = 0.644$, $p = 0.009$ on medical_abstracts; $\tau = 0.724$, $p < 0.001$ on mental_health) but much weaker agreement when LLM features are involved ($\tau = 0.289$, $p = 0.291$ for sentence transformer vs. LLM on medical_abstracts). This confirms that the LLM feature space captures fundamentally different structural information.

5.5 Ablation Studies

Systematic ablation across six dimensions on 3,238 instances establishes the robustness of our approach.

k-sensitivity. CRND noise detection AUC is stable across $k = 5$ to 50 (range < 0.04 on medical_abstracts), though mean CRND decreases monotonically with k (from 0.916 at $k = 5$ to 0.867 at $k = 50$). Figure 4 shows the sensitivity analysis.

Distance metric. Switching between Euclidean, cosine, and Manhattan distance changes CRND by at most 0.018 on any dataset, confirming metric robustness.

PCA dimensionality for Schoener's D. The 2D vs. 5D Schoener's D values correlate at $r = 0.55$ on average, indicating moderate stability with some information gain from additional dimensions.

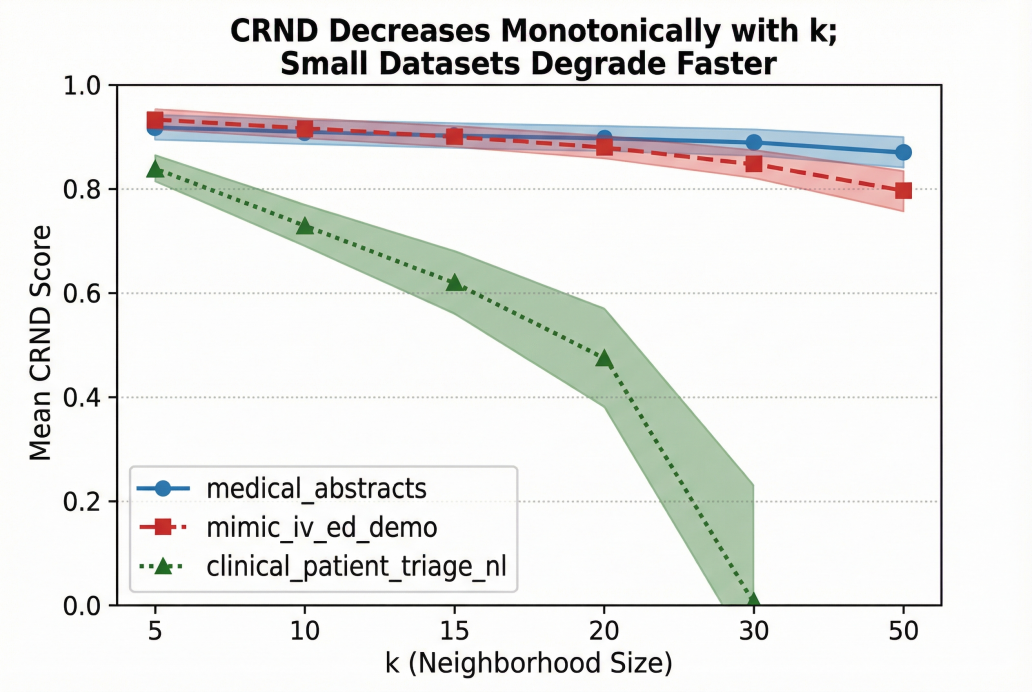

Figure 4: CRND sensitivity to neighborhood size k across three datasets. Large datasets (medical_abstracts, mimic_iv_ed) show gradual, monotonic decrease, while the small clinical_triage_nl dataset ($N = 31$) degrades sharply as k approaches N.

Alternative CRND formulations. Jaccard-based CRND performs best; Rank-Biased Overlap (RBO) with $p = 0.9$ and weighted Jaccard perform comparably. Pairwise decomposition reveals that the TF-IDF vs. LLM pair is the most informative.

Confound analysis. Partial Spearman correlations controlling for outlier score, boundary proximity, and vocabulary rarity show that the raw CRND–noise correlations (already near zero) become even closer to zero after controlling for confounds (partial ρ ranges from -0.049 to 0.020 across datasets).

6 Discussion

6.1 The Fundamental Failure of Cross-Representation Noise Detection

The decisive failure of CRND for noise detection (AUC = 0.497 vs. baseline 0.878) reveals a conceptual insight: cross-representation neighborhood instability and label noise are orthogonal signals. A mislabeled instance has inconsistent neighbors in every feature space—its k-NN neighborhood in TF-IDF space is label-confused, and so is its neighborhood in embedding space. The *consistency* of this confusion across spaces means that mislabeled instances do not exhibit higher cross-representation *dissonance* than correctly labeled instances. In contrast, baselines like kDN and cleanlab detect noise by identifying label inconsistency *within* each space—the correct level of analysis.

This finding has broader implications: methods that exploit inter-representation disagreement are not suitable for detecting phenomena (like label noise) that affect all representations similarly.

Inter-representation analysis is informative only when the phenomenon of interest manifests differently across representations—a condition satisfied by class structure (SC3) but not by random label corruption (SC1).

6.2 The Partial Promise of Ecological Niche Overlap

The adaptation of Schoener's D from ecology to ML class distributions represents a genuinely novel cross-domain transfer, confirmed across three independent literature surveys finding zero prior applications [Broennimann et al., 2012, Schoener, 1968, Warren et al., 2008]. The metric is mathematically equivalent to $1 - \mathrm{TV}(f, g)$ where TV is total variation distance, but its application within the Broennimann PCA-env framework—with PCA projection to 2D, grid-based KDE, and bandwidth selection via Scott's rule—provides a principled, reproducible pipeline for measuring class overlap.

The niche overlap profiles reveal substantial representation-dependent structure: the mean D-gap of 0.368 across 88 class pairs shows that the same class pair can be well-separated in one feature space (low D) while heavily overlapping in another (high D). However, this structural insight does not translate reliably into method selection prediction (pooled $\tau = 0.211$). The 73.6% drop in τ when replacing proxy features with true LLM features suggests that degraded LLM feature quality (one-hot encoding of text responses rather than proper logprob vectors) may partially explain the weak signal.

6.3 Class-Level CRND as a Descriptive Diagnostic

The success of SC3 ($\eta^2 = 0.162$) demonstrates that CRND captures meaningful class-level variation. In the mental_health dataset, "personality disorder" (mean CRND = 0.777) has the most stable neighborhood structure across representations, while "normal" (mean CRND = 0.962) shifts maximally. This likely reflects whether class boundaries are defined by specific lexical cues (stable across TF-IDF and embeddings) or by more abstract semantic relationships (representation-dependent).

The boundary stratification analysis (Cohen's $d = 0.565$) confirms that instances near class boundaries have elevated CRND, consistent with the interpretation that cross-representation instability is driven by class geometry rather than random variation.

6.4 Limitations

Several limitations constrain our conclusions. First, all noise detection experiments use synthetic noise injection (random label flips), which may not reflect the structure of real clinical mislabeling patterns. Second, LLM feature quality was compromised: OpenRouter did not provide logprobs for Llama-3.1-8B, forcing one-hot encoding of text responses. Third, two datasets are underpowered (MIMIC-IV-ED demo with 207 instances and clinical_triage_nl with only 31), affecting KDE reliability. Fourth, the PCA projection to 2D for Schoener's D may lose substantial variance from high-dimensional spaces. Fifth, the positive SC3 result could be confounded by class size differences or text length variation, which were not fully controlled.

7 Conclusion

We introduced Cross-Representation Neighborhood Dissonance (CRND) and adapted ecological niche overlap metrics (Schoener's D) for analyzing class structure across multiple feature spaces in clinical text classification. Our primary hypothesis—that cross-representation neighborhood

instability detects label noise—was decisively disconfirmed (AUC = 0.497 vs. 0.878). The method selection hypothesis received only weak, inconsistent support ($\tau = 0.211$). However, CRND reveals significant class-level variation in representation stability ($\eta^2 = 0.162$), and the ecological-to-ML transfer of Schoener's D is confirmed as genuinely novel and non-redundant with existing overlap measures.

We present this work as a transparent negative-result study, contributing an honest empirical account that we hope steers future work toward more productive applications of cross-representation analysis: characterizing class geometry, identifying representation-sensitive subgroups, and flagging instances where automated classification should defer to human judgment—while avoiding the tempting but unfounded assumption that cross-representation disagreement signals label quality.

Reproducibility. All experiments use CPU-only computation with publicly available models and datasets. Total LLM API cost across all experiments was under $0.05. Code and experimental outputs are available upon request.

References

A. S. Alqahtani et al. Clinical impact of artificial intelligence-based triage systems in emergency departments: A systematic review. *PMC*, 2025.

Dara Bahri, Heinrich Jiang, and Maya R. Gupta. Deep k-nn for noisy labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 540–550. PMLR, 2020.

Olivier Broennimann, Matthew C. Fitzpatrick, Peter B. Pearman, Blaise Petitpierre, Loïc Pellissier, Nigel G. Yoccoz, Wilfried Thuiller, Marie-Josée Fortin, Christophe Randin, Niklaus E. Zimmermann, Catherine H. Graham, and Antoine Guisan. Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, 21(4):481–497, 2012. doi: 10.1111/j.1466-8238.2011.00698.x.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

V. Fernandez-Arias et al. Automated triage classification in emergency services using Spanish clinical notes: A comparative analysis between ALBERT and classical machine learning approaches. In *NeurIPS 2025 Workshop*, 2025.

Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):289–300, 2002. doi: 10.1109/34.990132.

Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 4662–4671. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00463.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.

Ana Carolina Lorena, Luís Paulo F. Garcia, Jens Lehmann, Marcílio Carlos Pereira de Souto, and Tin Kam Ho. How complex is your classification problem?: A survey on measuring classification complexity. *ACM Comput. Surv.*, 52(5):107:1–107:34, 2019. doi: 10.1145/3347711.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Intell. Res.*, 70:1373–1411, 2021. doi: 10.1613/jair.1.12125.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, 2020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pages 6076–6085, 2017.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3980–3990. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

M. Rodriguez-Ruiz et al. Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review. *BMC Emergency Medicine*, 24, 2024.

Thomas W. Schoener. The Anolis lizards of Bimini: Resource partitioning in a complex fauna. *Ecology*, 49(4):704–726, 1968.

Michael R. Smith, Tony R. Martinez, and Christophe G. Giraud-Carrier. An instance level analysis of data complexity. *Mach. Learn.*, 95(2):225–256, 2014. doi: 10.1007/s10994-013-5422-z.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 9275–9293. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.746.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/arXiv.2302.13971.

Dan L. Warren, Richard E. Glor, and Michael Turelli. Environmental niche equivalency versus conservatism: Quantitative approaches to niche evolution. *Evolution*, 62(11):2868–2883, 2008. doi: 10.1111/j.1558-5646.2008.00482.x.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013.