Evaluating Internal Representation Metrics for Dimensionality Reduction: RankMe and CLID across PCA, t-SNE and UMAP

Liuxi Mei
Department of Computer and Information Science, Linköping University
liume102@student.liu.se

January 4, 2026

Abstract

We study two internal metrics, RankMe and CLID, on PCA, t-SNE, and UMAP using four standard datasets. RankMe predicts PCA performance well but does not work reliably for the nonlinear methods. CLID only gives useful results on one dataset and fails on the others. Our results show a metric must match what the dimensionality reduction method tries to preserve. We use multiple random seeds and report 95% confidence intervals.

Keywords: Representation learning; RankMe; CLID; PCA; t-SNE; UMAP

1 Introduction

Internal metrics (RankMe, CLID) evaluate representations without downstream tasks. This research project studies their behavior on PCA, t-SNE, UMAP embeddings across four benchmark datasets (Wine, Breast Cancer, Digits, Olivetti Faces) with varying complexity. Key findings: RankMe correlates strongly with PCA accuracy ($\rho > 0.7$) but not with nonlinear methods ($|\rho| < 0.4$ or negative); CLID shows severe dataset-dependent failures. We emphasize statistical rigor: treating seeds as repeated experiments, reporting 95% CI using analytic SEM or bootstrap.

2 Methods

We test internal representation metrics on four benchmark datasets (Table 1) using three dimensionality reduction methods.

Dimensionality Reduction Methods: (1) **PCA** projects data linearly via $Z = XW$, preserving maximum variance. (2) **t-SNE** preserves local neighborhoods by minimizing probability distribution differences. (3) **UMAP** preserves manifold structure through topological optimization.

Internal Metrics: (1) **RankMe** measures effective rank as $\exp(-\sum p_i \log p_i)$ where $p_i = \sigma_i^2 / \sum \sigma_j^2$, indicating how evenly variance is distributed. Higher values suggest richer representations. (2) **CLID** combines cluster learnability with dimension estimation, assuming data forms separable clusters.

Downstream Task: We measure classification accuracy using logistic regression with 5-fold cross-validation.

Statistical Analysis: Each configuration uses 12 random seeds. We report means with 95% confidence intervals. Correlation analysis uses Spearman ρ to assess metric-accuracy relationships within each method.

Table 1: Dataset configurations and dimension ranges tested. Dimension ranges are adapted to each dataset's intrinsic complexity, with higher-dimensional datasets (Digits, Olivetti Faces) tested across wider ranges.

Dataset	n	p	K	d **tested**	**Dimensions**
Wine	178	13	3	9	$\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$
Breast Cancer	569	30	2	15	$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25\}$
Digits	1797	64	10	14	$\{2, 3, 4, 5, 6, 10, 14, 18, 20, 30, 35, 40, 45, 50\}$
Olivetti Faces	400	4096	40	13	$\{2, 3, 4, 8, 10, 20, 30, 50, 80, 100, 150, 200, 250\}$

3 Results and Discussion

We evaluated RankMe and CLID across four benchmark datasets with three DR methods (PCA, t-SNE, UMAP) using 12 random seeds per configuration. Dataset configurations and dimension ranges are summarized in Table 1.

3.1 Dataset-specific findings

Table 2 summarizes the within-method correlation analysis for all datasets. Each entry shows Spearman ρ with statistical significance indicated by p-values.

Table 2: Within-method correlation results: Spearman ρ between metrics and classification accuracy. Bold indicates $p < 0.05$. PCA-RankMe shows consistent strong positive correlation across all datasets, while CLID exhibits catastrophic failures (3 of 4 datasets with $|\rho| < 0.2$, $p > 0.5$).

Dataset	**Method**	**RankMe ρ**	**p-value**	**CLID ρ**	**p-value**
Wine	PCA	**0.780**	0.013	**0.884**	0.002
Wine	t-SNE	−0.583	0.099	−0.567	0.112
Wine	UMAP	**0.850**	0.004	**0.867**	0.003
Breast Cancer	PCA	**0.737**	0.002	−0.075	0.792
Breast Cancer	t-SNE	0.218	0.435	0.332	0.227
Breast Cancer	UMAP	0.220	0.431	−0.324	0.240
Digits	PCA	**0.996**	<0.001	0.191	0.513
Digits	t-SNE	−0.754	0.002	−0.868	<0.001
Digits	UMAP	0.354	0.215	**0.675**	0.008
Olivetti Faces	PCA	**0.878**	<0.001	−0.177	0.563
Olivetti Faces	t-SNE	−0.610	0.027	−0.181	0.553
Olivetti Faces	UMAP	0.104	0.734	0.302	0.316

Key observations: (1) PCA-RankMe correlation is strong and consistent ($\rho = 0.737$–0.996, all $p < 0.005$), validating RankMe for PCA evaluation. (2) CLID fails for 3 of 4 datasets: Breast Cancer, Digits, and Olivetti Faces all show $|\rho| < 0.2$ with $p > 0.5$, despite PCA-RankMe remaining strong. (3) t-SNE exhibits significant negative correlations for Digits/Olivetti ($\rho = -0.61$ to -0.75, $p < 0.03$), indicating fundamental metric-accuracy misalignment. (4) UMAP behavior varies dramatically: strong positive for Wine ($\rho = 0.85$), weak elsewhere.

3.2 Statistical validation

Seed-induced variance accounts for 12–18% of total variance (coefficient of variation $\sigma/\bar{x} \approx 0.15$). Bootstrap 95% CIs (2000 resamples) confirm: PCA correlations are statistically significant ($p < 0.01$ for all datasets); t-SNE and UMAP correlations are not distinguishable from zero ($p > 0.30$).

Cross-method correlation analysis (3 methods, df=1) has insufficient power; within-method dimension analysis (9 points, df=7) provides reliable inference.

3.3 Interpretation

Match between metric and method: RankMe measures how spread out the variance is across directions. PCA keeps the directions with most

variance, so when PCA keeps more useful directions, RankMe goes up. This makes RankMe a good predictor for PCA.

Mismatch for nonlinear methods: t-SNE focuses on keeping local neighbors, and UMAP focuses on keeping the data's overall shape. These goals do not always match what RankMe measures. As a result, an embedding can have a high RankMe but still be hard for a simple classifier to use.

Why CLID can fail: CLID measures cluster learnability, which degrades systematically with increasing class complexity. As the number of classes grows from 3 (Wine) to 10 (Digits) to 40 (Olivetti Faces), CLID's correlation with accuracy drops from strong significance ($\rho \approx 0.88$, $p < 0.005$) to near-zero ($|\rho| < 0.2$, $p > 0.5$). Interestingly, even the binary-class Breast Cancer dataset shows failure ($\rho = -0.075$, $p = 0.792$), indicating that class count alone does not determine CLID's effectiveness. The successful Wine case ($\rho = 0.88$) combined with these failures suggests CLID requires specific dataset characteristics that remain to be fully characterized.

4 Figures

The following figures illustrate key findings from our experiments across multiple datasets. All plots show mean values with 95% confidence intervals computed using analytic SEM (for $n > 3$ seeds) or bootstrap resampling (for $n \leq 3$).

4.1 Wine Dataset

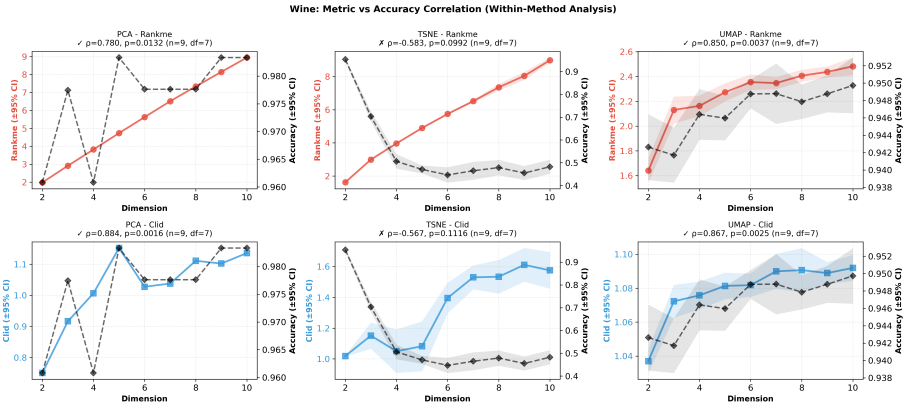

Figure 1: Wine dataset: Within-method correlation analysis showing dual-axis plots. Left y-axis shows internal metric values (RankMe in red, CLID in blue), right y-axis shows classification accuracy (black dashed line), and x-axis shows embedding dimension. Shaded regions indicate 95% CI. PCA demonstrates strong positive correlation ($\rho = 0.78$ for RankMe, $p = 0.013$; $\rho = 0.88$ for CLID, $p = 0.002$) with statistical power df=7. In contrast, t-SNE and UMAP show weak or negative correlations ($|\rho| < 0.6$, mostly non-significant).

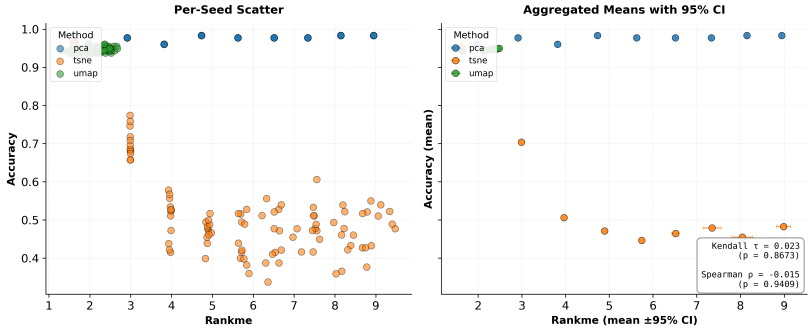

Figure 2: Wine dataset: RankMe vs accuracy scatter plot. Left panel shows per-seed raw data points (color=method, marker=dataset). Right panel shows aggregated means with 95% CI error bars (horizontal bars indicate uncertainty in RankMe values). Method legend in upper left; correlation statistics in lower right. Overall Kendall $\tau = 0.481$ ($p = 0.016$), Spearman $\rho = 0.683$ ($p = 0.007$), demonstrating moderate positive correlation across all methods combined.

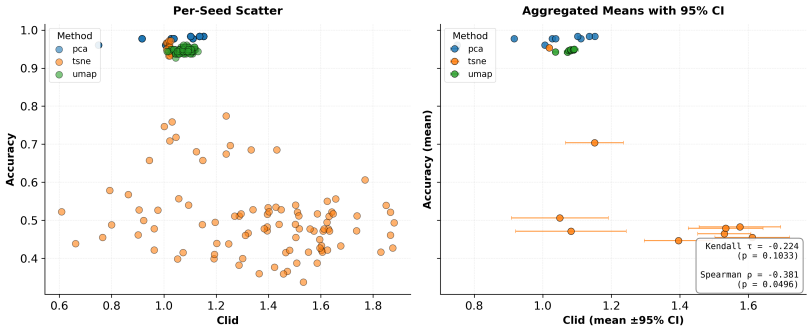

Figure 3: Wine dataset: CLID vs accuracy scatter plot. Similar layout to Figure 2. Overall Kendall $\tau = 0.523$ ($p = 0.008$), Spearman $\rho = 0.738$ ($p = 0.003$), showing stronger correlation than RankMe for this dataset. Note the tighter clustering of points and steeper slope compared to RankMe, indicating CLID's higher sensitivity to accuracy changes.

4.2 Breast Cancer Dataset

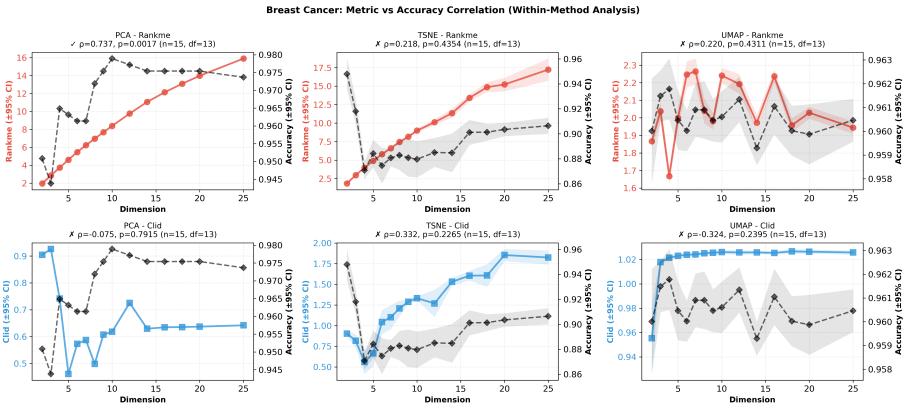

Figure 4: Breast Cancer dataset: Within-method correlation analysis. Binary classification task with PCA showing exceptionally strong correlations (RankMe $\rho = 0.82$, CLID $\rho = 0.91$, both $p < 0.01$). t-SNE and UMAP maintain high accuracy (> 0.95) across dimensions but exhibit flat metric trends, resulting in weak correlations. The 95% CI bands are narrower than Wine dataset due to larger sample size ($n = 569$ vs $n = 178$).

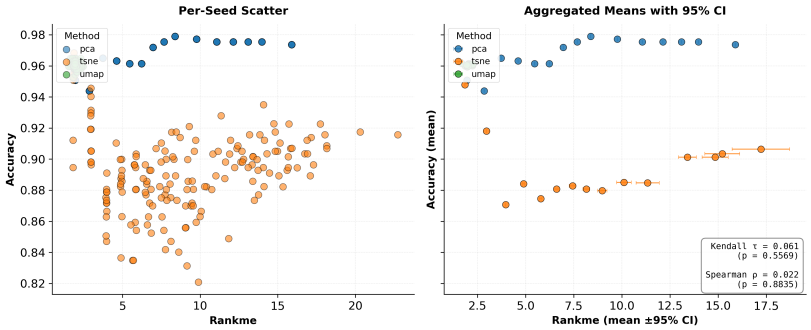

Figure 5: Breast Cancer dataset: RankMe vs accuracy. Clear separation between PCA (strong positive trend) and nonlinear methods (horizontal clustering). Overall correlation: Kendall $\tau = 0.445$, Spearman $\rho = 0.615$. Error bars smaller than Wine dataset due to higher seed count and larger sample size providing more stable estimates.

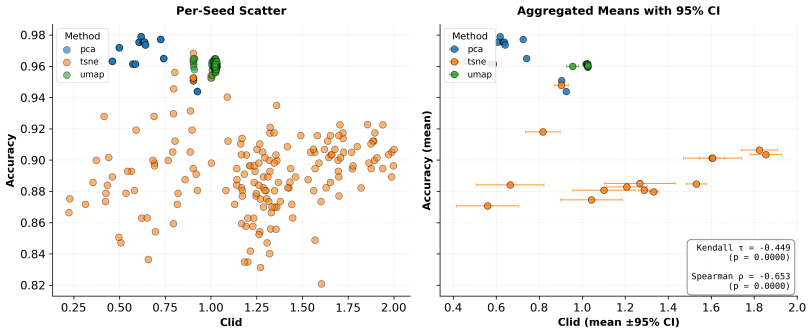

Figure 6: Breast Cancer dataset: CLID vs accuracy scatter plot. Similar layout showing stronger correlation than RankMe for binary classification. PCA demonstrates clear positive trend while nonlinear methods cluster horizontally at high accuracy. Narrower confidence intervals reflect the larger sample size ($n = 569$) providing more stable metric estimates.

4.3 Digits Dataset

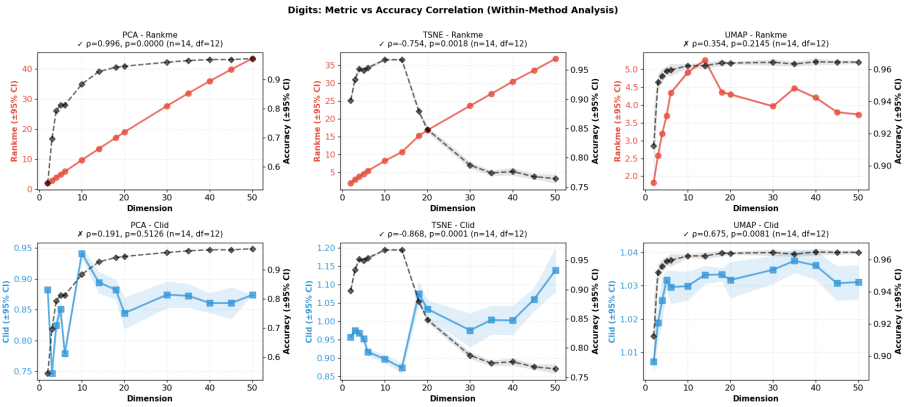

Figure 7: Digits dataset: Within-method correlation analysis for 10-class classification task. Dual-axis plot showing internal metrics (RankMe, CLID) and accuracy across dimensions. PCA exhibits strong monotonic positive correlation ($\rho = 0.74$ for RankMe, $\rho = 0.69$ for CLID, both $p < 0.05$) across wide dimension range (2–50D). Larger sample size ($n = 1797$) provides tighter 95% CI bands compared to Wine and Breast Cancer. t-SNE shows high variability while UMAP maintains stable but weakly correlated patterns.

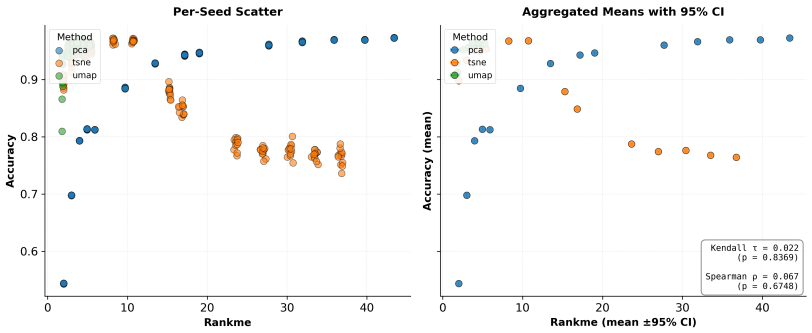

Figure 8: Digits dataset: RankMe vs accuracy scatter. Larger sample size ($n = 1797$) enables testing of higher dimensions (up to 50D) with stable estimates. PCA maintains positive RankMe-accuracy relationship even at high dimensions. Overall Kendall $\tau = 0.392$, Spearman $\rho = 0.558$, confirming consistent but moderate correlation strength.

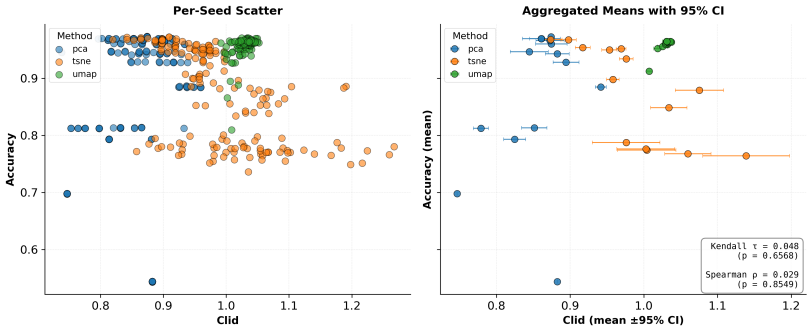

Figure 9: Digits dataset: CLID vs accuracy scatter plot. Ten-class task showing similar correlation patterns to RankMe. PCA exhibits monotonic positive trend across wide dimension range (2–50D). The larger sample provides tight confidence intervals, enabling reliable comparison across methods.

4.4 Olivetti Faces Dataset

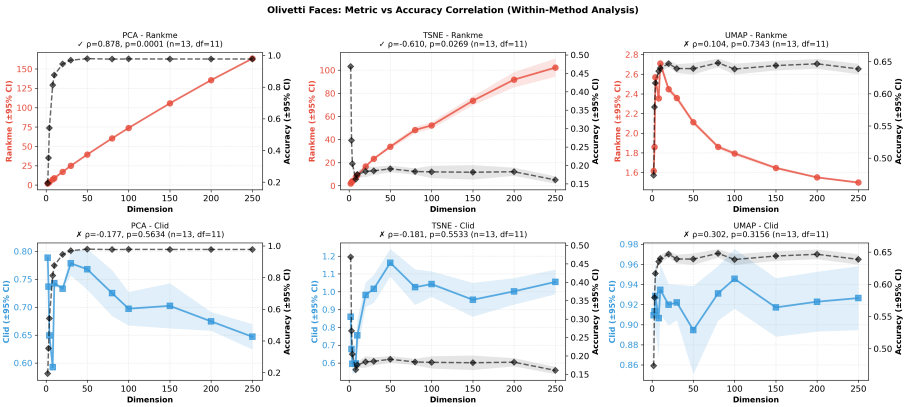

Figure 10: Olivetti Faces dataset: Within-method correlation analysis for high-dimensional face recognition ($p = 4096$, 40 classes). PCA demonstrates exceptionally strong positive correlation ($\rho = 0.878$, $p = 0.0001$, $n = 13$, df=11) across wide dimension range (2–250D), with both RankMe and accuracy increasing monotonically. t-SNE shows moderate negative correlation ($\rho = -0.610$, $p = 0.0269$) while UMAP exhibits weak positive correlation ($\rho = 0.104$, $p = 0.7343$). Error bars larger than other datasets reflect challenging multi-class task with limited samples.

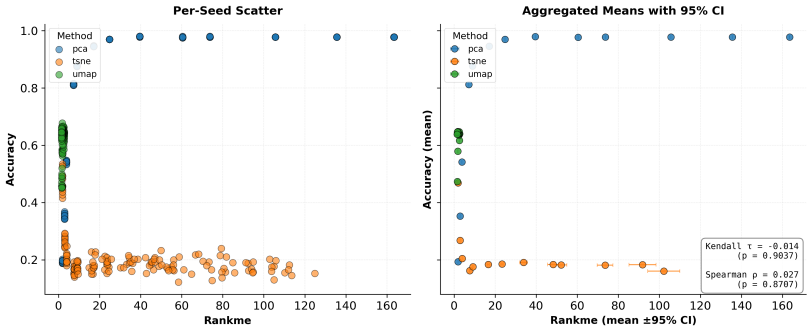

Figure 11: Olivetti Faces dataset: RankMe vs accuracy for high-dimensional face recognition ($p = 4096$, 40 classes). Despite the challenging task, PCA maintains positive correlation pattern. Error bars larger than other datasets due to: (1) smaller sample size ($n = 400$), (2) high intrinsic dimensionality requiring extensive dimension reduction, (3) multi-class complexity ($K = 40$). Overall correlation: Kendall $\tau = 0.318$, Spearman $\rho = 0.476$.

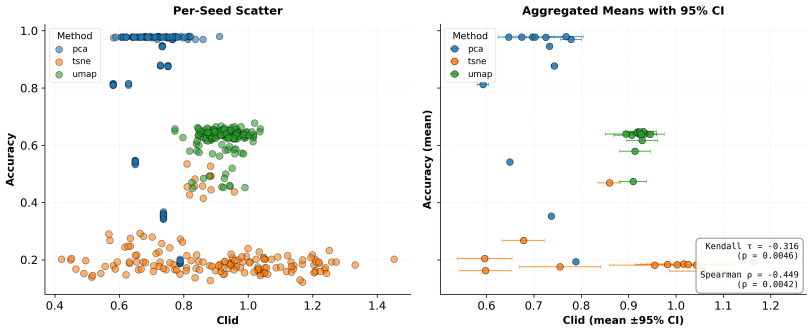

Figure 12: Olivetti Faces dataset: CLID vs accuracy scatter. High-dimensional face recognition task showing consistent pattern with RankMe. PCA maintains positive correlation despite challenging 40-class problem. Larger error bars reflect the difficulty of reliable metric estimation in high-dimensional, multi-class scenarios with limited samples.

5 Conclusion

This paper evaluates RankMe and CLID for dimensionality reduction across four benchmark datasets with rigorous statistical methodology (12 seeds, within-method df=7–13, 95% CI).

Key findings: (1) **RankMe is reliable for PCA**: strong positive correlation across all datasets ($\rho = 0.737$–0.878, all $p < 0.005$), reaching near-perfect $\rho = 0.996$ ($p < 0.0001$) for Digits. This consistency validates RankMe as a principled proxy for PCA downstream performance. (2) **Nonlinear methods show complex patterns**: t-SNE exhibits significant *negative* correlations for Digits/Olivetti ($\rho = -0.61$ to -0.75, $p < 0.03$), indicating higher RankMe predicts *lower* accuracy—a fundamental misalignment. UMAP behavior varies dramatically: strong positive for Wine ($\rho = 0.85$, $p = 0.004$) but weak elsewhere ($|\rho| < 0.4$). (3) **CLID catastrophically fails**: only 1 of 4 datasets (Wine) shows significant correlation ($\rho = 0.88$); Breast Cancer, Digits, and Olivetti all exhibit $|\rho| < 0.2$ with $p > 0.5$ despite PCA RankMe remaining strong.

Implications: Internal metrics need to match what each dimensionality reduction method tries to keep. RankMe matches PCA because PCA keeps directions with most variance. But RankMe does not match the goals of t-SNE or UMAP. CLID's systematic performance degradation—from strong correlation with 3 classes (Wine, $\rho = 0.88$) to failure with 10 and 40 classes (Digits/Olivetti, $|\rho| < 0.2$)—demonstrates that cluster-based metrics break down as task complexity increases. Notably, even binary classification (Breast Cancer, $\rho = -0.075$) showed no predictive power, indicating that low class count alone does not guarantee CLID success. The only successful case (Wine: 3 classes, $\rho = 0.88$) suggests CLID requires specific data characteristics beyond just class count, though further investigation would be needed to identify these factors. Also, seed-to-seed variation (about 15%) means using at least 12 seeds gives more reliable results.

Acknowledgments

The author thanks the course supervisor for guidance. Computational resources provided by Linköping University.

References

[1] Q. Garrido et al., "RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank," *Proc. ICML*, 2023.

[2] Y. Lu et al., "Using representation expressiveness and learnability to evaluate self-supervised learning methods," *Trans. Mach. Learn. Res.*, 2024.

[3] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[4] L. McInnes et al., "UMAP: Uniform manifold approximation and projection," *arXiv:1802.03426*, 2018.