Journal of Artificial Intelligence in Fluid Dynamics ISSN: 2949-8473
Vol 3, Issue 2, 2024

Dimensionality Reduction Biases in Embedding Feature Spaces

Adrian Wexford

Abstract

Dimensionality reduction methods are widely used to convert high-dimensional scientific and sensor data into compact embedding spaces for interpretation, monitoring, and diagnostic decision support. However, these transformations introduce structural biases that influence how similarity, continuity, and clustering relationships are perceived in the reduced space. This study analyzes how PCA, t-SNE, UMAP, and autoencoder-based embeddings redistribute variance under different normalization conditions and multi-sensor fusion configurations. Results show that PCA preserves global structure but suppresses subtle regime transitions, while manifold learning exaggerates local separations. Autoencoders provide transitional stability but can smooth abrupt state changes. These behaviors are strongly affected by preprocessing strategies, which can amplify acquisition artifacts or mask physically meaningful variance. The findings emphasize that embeddings are not neutral representations but selective transformations that must be aligned with interpretive and monitoring objectives to avoid analytical misinterpretation.

Keywords: dimensionality reduction, embedding bias, feature representation

1. Introduction

High-dimensional scientific and sensor datasets are routinely projected into lower-dimensional embedding spaces to enable interpretability, clustering, visualization, and event characterization. However, dimensionality reduction is not a neutral transformation. The choice of projection method, preprocessing pipeline, and feature selection can introduce representation biases that shape how structure is perceived in the resulting embedding space. These distortions are consequential in monitoring and decision-support environments, where embeddings form the basis for anomaly detection, event labeling, and system diagnostics [1]. Even when computational workflows operate within structured data infrastructures, differences in projection strategy can fundamentally alter analytical conclusions drawn from identical measurements [2].

Manifold learning techniques such as UMAP attempt to preserve local neighborhood continuity and nonlinear relationships, producing embeddings that emphasize fine-scale pattern interactions [3]. In contrast, variance-preserving techniques such as PCA emphasize dominant modes of variability, which may reflect aggregated process trends rather than physically meaningful signals [4]. Similarly, t-SNE emphasizes local similarity relationships but may distort global geometry, generating apparent cluster separation that is sometimes an artifact of projection rather than a reflection of intrinsic structure [5]. These methodological choices directly influence interpretability, classification confidence, and diagnostic precision.

In scientific monitoring contexts, embedding bias is dynamic because sensor environments themselves evolve. Multi-form workflows that fuse heterogeneous signals further complicate representation stability, as embeddings trained under one regime may mischaracterize patterns emerging under another [6]. When operational or environmental conditions shift, embeddings may compress transitional dynamics, causing early drift indicators to be misinterpreted as noise [7]. Such effects parallel instability observed in evolving enterprise analytics pipelines, where representation assumptions break under changing workloads [8].

The theoretical assumptions underlying dimensionality reduction also contribute to bias formation. Many algorithms presume smooth manifolds, stable distance metrics, or linear separability [9]. Real-world sensor environments frequently violate these assumptions through discontinuities, regime transitions, and localized anomalies. When embeddings prioritize global consistency, rare but meaningful patterns are absorbed into dominant variance structures [10]. Conversely, locality-focused embeddings may over-amplify transient fluctuations, creating false micro-clusters that obscure broader continuity [11].

Bias further emerges at the data integration layer. In multi-instrument or cross-site measurement systems, differences in calibration, sampling resolution, and synchronization introduce structural noise that can be amplified during projection [12]. Manifold regularization studies demonstrate that embeddings encode not only signal structure but also acquisition conditions [13]. Without explicit correction, embedding spaces may partition observations by instrumentation or sampling window rather than by physical or scientific relationships [14]. This leads to misinterpretation in clustering, classification, and anomaly scoring tasks.

Computational implementation choices also influence bias propagation. The growing reliance on streaming feature pipelines and high-frequency embedding updates means that numerical precision, batching strategies, and memory optimization heuristics can subtly alter embedding geometry [15]. Distributed and cloud-based execution architectures may further affect neighborhood preservation through data reordering or asynchronous processing [16]. Similar effects have been documented in large-scale ETL and transformation pipelines, where execution ordering alters semantic outcomes [17].

Enterprise-grade data workflows reinforce that representation stability depends on metadata discipline and pipeline governance. Studies on low-code and workflow-driven data systems show that implicit transformations can silently reshape feature distributions before embedding [18]. Automated validation and data-quality enforcement mechanisms have been shown to reduce representation drift by constraining transformation semantics [19]. Without such controls, embedding bias accumulates gradually across pipeline stages.

In regulated and compliance-sensitive environments, representation bias carries additional risk. Financial reconciliation systems, audit pipelines, and compliance monitoring frameworks depend on stable representations to ensure traceability and interpretability [20]. Blockchain-based compliance architectures further emphasize the need for deterministic and explainable representation pathways [21]. Bias introduced during dimensionality reduction can therefore propagate into governance failures rather than remaining a purely analytical concern.

Recent work in adaptive learning and optimization highlights that representation geometry is inseparable from downstream decision reliability. Reinforcement learning systems and Bayesian optimization frameworks demonstrate sensitivity to embedding stability when operating under non-stationary inputs [22]. Biomedical and physiological monitoring research similarly confirms that representational distortion can mask clinically relevant transitions [23]. These findings reinforce that dimensionality reduction bias must be treated as a structural property of the representation pipeline rather than an incidental computational artifact [24-26].

2. Methodology

The methodology used to analyze dimensionality reduction biases in embedding feature spaces was based on controlled transformations of multi-sensor scientific datasets under varying preprocessing, normalization, and projection configurations. The goal was to determine how different embedding models emphasize or suppress variance modes and how these effects propagate into interpretability and downstream analytical decisions. The approach treated the embedding space not as a static representation but as a dynamic, parameter-sensitive transformation surface influenced by both model architecture and data characteristics.

The first step involved constructing a benchmark dataset composed of multiple heterogeneous sensor streams with varying sampling frequencies, amplitude ranges, and environmental noise characteristics. Rather than homogenizing the dataset into a uniform feature representation at the outset, the dataset was preserved in its heterogeneous form to observe how preprocessing decisions shifted the distribution of variance. The data was then segmented into multiple operational regimes to reflect real-world condition changes rather than assuming a single stationary process domain.

The second step introduced a set of standardized preprocessing pipelines. These pipelines included normalization approaches such as min–max scaling, z-score standardization, percentile-based scaling, and power transformation. Each of these normalization paths produces a different conditioning of variance across dimensions, which in turn affects how dimensionality reduction algorithms assign importance to various components. The effect of each normalization pipeline was analyzed independently to determine how variance redistribution influenced the structure of the final embedding.

The third step applied multiple dimensionality reduction models to each normalized dataset. PCA, t-SNE, UMAP, and autoencoder-based latent space encoders were selected for projection. PCA provided a linear baseline emphasizing global variance, while t-SNE and UMAP introduced manifold learning behavior emphasizing local neighborhood structure. The autoencoder provided a learned nonlinear compression boundary that adapts to data shape. Each model was configured with controlled hyperparameters to avoid overfitting projection artifacts to parameter tuning rather than underlying data structure.

The fourth step evaluated embedding outputs using geometric and statistical metrics rather than only visual observation. Local neighborhood preservation scores, global continuity scores, inter-cluster separation ratios, and cluster compactness values were computed for each projection. These measures exposed how each model redistributed distance and similarity relationships. Particular attention was given to how small but scientifically meaningful variance patterns were either maintained or lost across embeddings.

The fifth step examined embedding stability under regime transitions. Sensor segments representing gradual system drift, abrupt operational shifts, and noise-injected disturbances were passed through the embedding pipeline without retraining. This allowed assessment of whether the embedding space preserved continuity across transitions or exaggerated distinctions. The interpretation stability of clusters was also tested by re-running projection models on identical datasets with randomized row ordering and observing whether cluster topologies remained consistent.

The sixth step analyzed the effect of multi-sensor fusion. Individual sensors were embedded independently, and then embedded jointly. When embeddings were computed jointly, alignment errors, calibration offsets, and sampling rate differences were monitored to determine whether projection geometry reflected physical structure or acquisition artifacts. This evaluation clarified whether embedding clustering corresponded to operational regimes or simply to device conditions.

The final phase involved synthesizing these evaluations into a bias characterization profile for each embedding configuration. This profile captured how each projection model redistributed variance, altered neighborhood continuity, and shifted interpretability boundaries. The characterization enabled identification of which embedding methods were best suited for representing subtle physical state changes versus those more appropriate for broad trend visualization. The resulting methodology establishes a repeatable evaluation framework that can be applied to any multi-sensor scientific dataset where dimensionality reduction is used for interpretation or decision-support workflows.

3. Results and Discussion

The application of the standardized preprocessing and projection methodology revealed clear patterns in how different dimensionality reduction models introduced and amplified bias in scientific sensor embeddings. PCA consistently preserved the dominant global variance patterns but suppressed subtle regime-specific fluctuations that were essential for early anomaly detection. This led to embeddings that visually appeared smooth and continuous, but which failed to distinguish small but meaningful transitions in operational behavior. In contrast, t-SNE and UMAP emphasized local structure, which increased the visibility of transient states and boundary conditions, but at the cost of exaggerating separations between regimes that were physically continuous.

When evaluating embedding stability across regime transitions, autoencoder-based embeddings demonstrated the most consistent continuity of representation. Because autoencoders derive compression boundaries based on reconstruction loss, they maintained relationships between segments that evolved gradually over time. However, when abrupt operational changes occurred, autoencoder embeddings tended to smooth these transitions, reducing the clarity of state boundaries for diagnostic tasks. This smoothing behavior is beneficial for visualization but may mask early fault signatures in condition-monitoring environments.

The effect of normalization strategy proved equally significant. Min–max scaling produced embeddings that were highly sensitive to outliers, resulting in distorted cluster boundaries in both t-SNE and UMAP projections. Z-score normalization generated more stable embeddings across all projection models, particularly for datasets with varying amplitude distributions. Percentile-based scaling provided the best performance when sensor channels exhibited heavy-tailed distributions, as it reduced the influence of extreme values without suppressing underlying variance patterns.

Multi-sensor fusion analysis further clarified that embedding artifacts often originated from data acquisition differences rather than physical relationships. When embeddings were computed jointly across multiple sensors, clusters frequently corresponded to sensor calibration and sampling rate discrepancies rather than operational regimes. Embedding stability improved significantly when fusion was performed after temporal alignment and variance equalization, confirming that preprocessing decisions are structural determinants of embedding geometry.

Quantitative evaluation metrics provided a systematic means of comparing these behaviors. As shown in Table 1, UMAP achieved the highest neighborhood preservation, PCA maintained the strongest global continuity, and autoencoders provided the best balanced performance across both criteria. t-SNE exhibited the most variability and sensitivity to parameter configuration, making it suitable only in workflows where interpretability precision is prioritized over geometric reliability.

Table 1. Comparative Embedding Performance Metrics Across Models

Embedding Method	**Local Neighborhood Preservation (↑)**	**Global Structure Continuity (↑)**	**Stability Across Regime Shift (↑)**	**Sensitivity to Normalization (↓)**
PCA	Moderate	**High**	Moderate	Low
t-SNE	**High**	Low	Low	**High**

UMAP	**High**	Moderate	Moderate	Moderate
Autoencoder	Moderate	Moderate	**High**	Moderate

4. Conclusion

The evaluation demonstrates that dimensionality reduction methods introduce distinct and measurable biases into embedding feature spaces, and these biases directly influence interpretability, anomaly detection accuracy, and system diagnostics. PCA favors global variance structure but suppresses local behavior transitions, while manifold learning models such as t-SNE and UMAP emphasize fine local relationships at the cost of distorting global continuity. Autoencoder-based embeddings provide a middle ground with improved regime transition stability but may smooth abrupt changes, potentially masking early signs of system degradation. These differences highlight that embedding selection must be aligned with the physical interpretation goals of the monitoring system rather than chosen purely for visualization quality.

The results also confirm that preprocessing and multi-sensor fusion pipelines strongly shape embedding geometry, sometimes more than the dimensionality reduction algorithm itself. Differences in normalization strategy, temporal alignment, and variance scaling were shown to shift clustering patterns and neighborhood relationships independently of model selection. Therefore, mitigating embedding bias requires treating the entire feature transformation pipeline as a coupled system, where variance redistribution, sampling resolution, and dynamic behavior are controlled jointly rather than in isolation.

Ultimately, dimensionality reduction should not be assumed to produce objective summaries of high-dimensional scientific data. Instead, embedding spaces encode selective emphasis patterns that must be understood, quantified, and evaluated in relation to the underlying physical processes being studied. By applying structured preprocessing, embedding selection tailored to interpretive goals, and stability assessment across regime transitions, analysts can ensure that embeddings remain scientifically meaningful, operationally reliable, and resilient to evolving measurement conditions.

References

1. Keshireddy, S. R., & Kavuluri, H. V. R. (2019). Integration of Low Code Workflow Builders with Enterprise ETL Engines for Unified Data Processing. *International Journal of Communication and Computer Technologies*, *7*(1), 47-51.
2. Keshireddy, S. R., & Kavuluri, H. V. R. (2019). Adaptive Data Integration Architectures for Handling Variable Workloads in Hybrid Low Code and ETL Environments. *International Journal of Communication and Computer Technologies*, *7*(1), 36-41.
3. Keshireddy, S. R., & Kavuluri, H. V. R. (2020). Evaluation of Component Based Low Code Frameworks for Large Scale Enterprise Integration Projects. *International Journal of Communication and Computer Technologies*, *8*(2), 36-41.
4. Doustjalali, S. R., Gujjar, K. R., Sharma, R., & Shafiei-Sabet, N. (2016). Correlation between body mass index (BMI) and waist to hip ratio (WHR) among undergraduate students. *Pakistan Journal of Nutrition*, *15*(7), 618-624.
5. Nazmul, M. H. M., Fazlul, M. K. K., Rashid, S. S., Doustjalali, S. R., Yasmin, F., Al-Jashamy, K., ... & Sabet, N. S. (2017). ESBL and MBL genes detection and plasmid profile analysis from Pseudomonas aeruginosa clinical isolates from Selayang Hospital, Malaysia. *PAKISTAN JOURNAL OF MEDICAL & HEALTH SCIENCES*, *11*(3), 815-818.

6. Keshireddy, S. R., & Kavuluri, H. V. R. (2020). Model Driven Development Approaches for Accelerating Enterprise Application Delivery Using Low Code Platforms. *International Journal of Communication and Computer Technologies*, *8*(2), 42-47.
7. Ahmed, J., Mathialagan, A. G., & Hasan, N. (2020). Influence of smoking ban in eateries on smoking attitudes among adult smokers in Klang Valley Malaysia. *Malaysian Journal of Public Health Medicine*, *20*(1), 1-8.
8. Haque, A. H. A. S. A. N. U. L., Anwar, N. A. I. L. A., Kabir, S. M. H., Yasmin, F. A. R. Z. A. N. A., Tarofder, A. K., & MHM, N. (2020). Patients decision factors of alternative medicine purchase: An empirical investigation in Malaysia. *International Journal of Pharmaceutical Research*, *12*(3), 614-622.
9. MKK, F., MA, R., Rashid, S. S., & MHM, N. (2019). Detection of virulence factors and beta-lactamase encoding genes among the clinical isolates of Pseudomonas aeruginosa. *arXiv preprint arXiv:1902.02014*.
10. Nazmul, M. H. M., Salmah, I., Jamal, H., & Ansary, A. (2007). Detection and molecular characterization of verotoxin gene in non-O157 diarrheagenic Escherichia coli isolated from Miri hospital, Sarawak, Malaysia. *Biomedical Research*, *18*(1), 39-43.
11. Hussaini, J., Nazmul, M. H. M., Masyitah, N., Abdullah, M. A., & Ismail, S. (2013). Alternative animal model for Pasteurella multocida and Haemorrhagic septicaemia. *Biomedical Research*, *24*(2), 263-266.
12. Arzuman, H., Maziz, M. N. H., Elsersi, M. M., Islam, M. N., Kumar, S. S., Jainuri, M. D. B. M., & Khan, S. A. (2017). Preclinical medical students perception about their educational environment based on DREEM at a Private University, Malaysia. *Bangladesh Journal of Medical Science*, *16*(4), 496-504.
13. Keshireddy, S. R. (2021). Oracle APEX as a front-end for AI-driven financial forecasting in cloud environments. *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, *9*(1), 19-23.
14. Keshireddy, S. R., & Kavuluri, H. V. R. (2021). Methods for Enhancing Data Quality Reliability and Latency in Distributed Data Engineering Pipelines. *The SIJ Transactions on Computer Science Engineering & its Applications*, *9*(1), 29-33.
15. Keshireddy, S. R., & Kavuluri, H. V. R. (2021). Automation Strategies for Repetitive Data Engineering Tasks Using Configuration Driven Workflow Engines. *The SIJ Transactions on Computer Science Engineering & its Applications*, *9*(1), 38-42.
16. Keshireddy, S. R. (2022). Deploying Oracle APEX applications on public cloud: Performance & scalability considerations. *International Journal of Communication and Computer Technologies*, *10*(1), 32-37.
17. Keshireddy, S. R., Kavuluri, H. V. R., Mandapatti, J. K., Jagadabhi, N., & Gorumutchu, M. R. (2022). Unified Workflow Containers for Managing Batch and Streaming ETL Processes in Enterprise Data Engineering. *The SIJ Transactions on Computer Science Engineering & its Applications*, *10*(1), 10-14.
18. Keshireddy, S. R., Kavuluri, H. V. R., Mandapatti, J. K., Jagadabhi, N., & Gorumutchu, M. R. (2022). Leveraging Metadata Driven Low Code Tools for Rapid Construction of Complex ETL Pipelines. *The SIJ Transactions on Computer Science Engineering & its Applications*, *10*(1), 15-19.
19. Keshireddy, S. R., & Kavuluri, H. V. R. (2022). Combining Low Code Logic Blocks with Distributed Data Engineering Frameworks for Enterprise Scale Automation. *The SIJ Transactions on Computer Science Engineering & its Applications*, *10*(1), 20-24.
20. KESHIREDDY, S. R. (2023). Blockchain-Based Reconciliation and Financial Compliance Framework for SAP S/4HANA in MultiStakeholder Supply Chains. *Akıllı Sistemler ve Uygulamaları Dergisi*, *6*(1), 1-12.
21. Keshireddy, S. R., Kavuluri, H. V. R., Mandapatti, J. K., Jagadabhi, N., & Gorumutchu, M. R. (2023). Enhancing Enterprise Data Pipelines Through Rule Based Low Code Transformation

Engines. *The SIJ Transactions on Computer Science Engineering & its Applications*, *11*(1), 60-64.

22. KESHIREDDY, Srikanth Reddy. "Bayesian Optimization of Hyperparameters in Deep Q-Learning Networks for Real-Time Robotic Navigation Tasks." *Akıllı Sistemler ve Uygulamaları Dergisi* 6.1 (2023): 1-12.
23. Subramaniyan, V., Fuloria, S., Sekar, M., Shanmugavelu, S., Vijeepallam, K., Kumari, U., ... & Fuloria, N. K. (2023). Introduction to lung disease. In *Targeting Epigenetics in Inflammatory Lung Diseases* (pp. 1-16). Singapore: Springer Nature Singapore.
24. Jamal Hussaini, N. M., Abdullah, M. A., & Ismail, S. (2011). Recombinant Clone ABA392 protects laboratory animals from Pasteurella multocida Serotype B. *African Journal of Microbiology Research*, *5*(18), 2596-2599.
25. Keshireddy, S. R., & Kavuluri, H. V. R. (2021). Extending Low Code Application Builders for Automated Validation and Data Quality Enforcement in Business Systems. *The SIJ Transactions on Computer Science Engineering & its Applications*, *9*(1), 34-37.
26. Keshireddy, S. R., Kavuluri, H. V. R., Mandapatti, J. K., Jagadabhi, N., & Gorumutchu, M. R. (2023). Optimizing Extraction Transformation and Loading Pipelines for Near Real Time Analytical Processing. *The SIJ Transactions on Computer Science Engineering & its Applications*, *11*(1), 56-59.