JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH001107

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Rethinking Collinearity in Self-Organizing Maps: Evidence From Geophysical Data Classification

Limin Xu[1], **Leonardo Feltrin**[2], **and Eleanor C. R. Green**[1]

[1]School of Geography, Earth and Atmospheric Sciences, University of Melbourne, Melbourne, VIC, Australia, [2]Mineral Intelligence Group, Mineral Economy Solutions Unit, Geological Survey of Finland (GTK), Espoo, Finland

Abstract This study examines the impact of collinearity on unsupervised machine learning algorithms (UMLAs), specifically Self-Organizing Maps (SOMs), for detecting lithological boundaries in geophysical data. Using a multi-scale experimental framework that includes bivariate isotropic clusters, geologically complex Noddy simulations, and real-world data from Victoria, Australia, we show that SOMs exhibit inherent robustness to collinearity in variable geological settings due to their distance-based optimization and neighborhood smoothing. Performance evaluation across UMLAs reveals that distance-based algorithms maintain stability under collinear conditions, while other methodologies such as Agglomerative Clustering show sensitivity or K-means and Gaussian Mixtures Models (GMMs) show classification performance degradation. Critically, collinear features improve classification when cluster separation is minimal in synthetic models, and derivative transforms enhance boundary detection, despite high predictor correlation. We propose the use of a Geological Complexity Index (GCI) analysis to identify areas prone to collinearity issues by geospatially mapping a novel distinction between geologically meaningful, model-relevant "good" collinearity in high-GCI settings and redundancy-related "bad" collinearity in low-GCI regions, which are typically characterized by high noise-to-signal ratios.

Plain Language Summary Self-Organizing Maps (SOMs) perform reliably when analyzing correlated geophysical measurements, unlike traditional statistical clustering methods that struggle with redundant data. This robustness stems from their distance-based learning approach, which naturally handles correlations between features. When geological boundaries overlap or are subtle, including correlated features can actually improve classification accuracy rather than degrade it. We introduce a method to distinguish between beneficial correlations—those reflecting genuine geological relationships in complex settings—and problematic redundancy in simpler, noise-dominated environments. This distinction helps practitioners decide when correlated features enhance model performance and when they should be removed.

Key Points:

- SOMs show robustness to collinearity via distance-based optimization, maintaining stable performance across varying geological complexity
- Collinear features enhance classification when cluster separation is minimal; derivative transforms improve boundary detection accuracy
- The GCI framework spatially distinguishes beneficial geological collinearity in complex regions from problematic redundancy in noise-dominated areas

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

L. Xu,
liminx1@student.unimelb.edu.au

Citation:

Xu, L., Feltrin, L., & Green, E. C. R. (2026). Rethinking collinearity in self-organizing maps: Evidence from geophysical data classification. *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2025JH001107. https://doi.org/10.1029/2025JH001107

Received 31 OCT 2025
Accepted 12 FEB 2026

1. Introduction

Self-Organizing Maps (SOMs) have gained popularity as an unsupervised machine learning clustering algorithm (UMLA), which projects high-dimensional data onto a lower-dimensional space (Kohonen, 1990, 2013; Kohonen et al., 1996). They can be used for pattern recognition and classification in geophysical data analysis, such as identifying geological boundaries (Carneiro et al., 2012) and classifying lithological units from complex, multivariate geophysical data sets (Cracknell & Reading, 2014; Reading et al., 2015; Xu et al., 2025b).

A common concern in multivariate analysis is collinearity (Tomaschek et al., 2018), where predictor variables are highly correlated with each other (Morlini, 2006). In traditional statistical methods, such as multiple linear regression (Morrissey & Ruxton, 2018), collinearity leads to singularity problems during matrix inversion, resulting in unstable parameter estimates and reduced predictive power (Mason & Perreault, 1991). Consequently, preprocessing techniques, such as principal component analysis (PCA), are often applied to eliminate redundancy before analysis (Dascălu & Cozma, 2009; Jolliffe, 2002). However, there is growing evidence that UMLAs, including SOMs, may respond differently to collinearity compared to classical statistical methods (Curtis & Ghosh, 2011; Palomino-Echeverria et al., 2024).

This distinction is particularly relevant in geophysical applications, where derived data sets often exhibit spatially variable collinearity due to their shared origin and underlying geological complexity (Brazell et al., 2019; Feyen & Caers, 2006; Fouedjio & Klump, 2019). Geophysical data processing commonly involves calculating various transforms from raw measurements to enhance specific features of interest (Dentith & Mudge, 2014;

Fairhead, 2015). Common transforms include vertical derivative (1VD), which enhances shallow features and sharpens boundaries; tilt angle, which normalizes signal amplitude and enhances edges regardless of depth; and analytical signal amplitude (ASA), which responds to magnetization boundaries regardless of direction (Isles & Rankin, 2013; Jacobsen, 1987; Khalil et al., 2016; Smith et al., 2022). Previous studies have shown benefits from combining multiple transforms in geological interpretation (Cooper & Cowan, 2008; Holden et al., 2008). Each transform emphasizes different aspects of the underlying geological structures, potentially providing complementary information (Xu et al., 2024; Xu & Green, 2023). Nevertheless, these transforms are mathematically derived from the same raw measurements and frequently exhibit high mutual correlation. In aeromagnetic surveys —commonly deployed for mineral exploration—transforms including ASA and 1VD show correlation coefficients exceeding 0.8 in homogeneous lithological domains (Doo et al., 2007; Reid, 2007; Reid et al., 1990). Similarly, gravity gradiometry transforms, for example, tensor components, display domain-specific collinearity patterns due to shared source-field relationships (Cao et al., 2024). Recent advances in computational geophysics and hydrology have examined collinearity variations across spatial domains using moving-window analysis approaches and data assimilation. For example, ensemble smoother methods with covariance localization explicitly address domain-dependent correlations to reduce spurious dependencies (Todaro et al., 2021). In geophysics, Euler deconvolution workflows employ sliding-window analysis to improve solution stability and depth estimation reliability by reducing the effects of noise and incorrect structural index selection (Reid et al., 1990; Stavrev & Reid, 2007; Thompson, 1982; Uieda & Barbosa, 2012). This methodology is similar to part of the workflow adopted in fractal complexity analysis used in mineralization characterization studies (Ford & Blenkinsop, 2008). It allows quantification of local statistical relationships between geophysical parameters.

The proposed multi-scale experimental framework expands further on sliding window approaches allowing a closer (three-dimensional) evaluation of the impact of collinearity on the performance of UMLAs, particularly SOMs, for 3D lithological boundary detection. Progressing from controlled synthetic models to real-world scenarios in Victoria, Australia, we also consider the transition from synthetic (less geologically complex) models to more realistic scenarios. Central to this work is the use of forward modeling as a controlled laboratory for collinearity analysis. By generating synthetic geological models with defined architectures and simulating their geophysical responses using Noddy (Guo et al., 2021; Jessell & Valenta, 1996), we enable the first systematic investigation of how geological structures manifest as distinct patterns of collinearity in data. This approach uniquely allows the characterization of geologically meaningful collinearity versus statistically redundant collinearity, moving beyond traditional statistical treatments. The framework progresses from controlled synthetic experiments to a real-world case study in Victoria, Australia (Xu et al., 2025b), addressing three research questions: (a) How does spatial variability in collinearity affect pattern classification across geological contexts? (b) Can regional/areal collinearity mapping inform local algorithm selection and tuning pipelines to improve geophysical processing and in turn facilitate geological interpretation? (c) What key differences emerge between synthetic and real-world classification outcomes, and how can they guide future machine learning applications in geophysics? Quantitative assessments of algorithm performance provide theoretical insights into collinearity sensitivity. Findings are contextualized by the scale of synthetic models and the scope of real-world validation, which may affect their applicability to other crustal settings, particularly those in different tectonic environments or in geologically and temporally distinct terranes.

2. Method

We implemented a three-stage experimental framework to examine collinearity effects in SOM-based geophysical data classification. Each stage shares a common methodological core: (1) generation of input features with known or quantifiable collinearity patterns, (2) application of SOM, and (3) quantitative evaluation against reference labels. The stages differ in data complexity and evaluation metrics: Stage 1 used isotropic clusters with induced statistical collinearity and implement auxiliary UMLAs, including K-Means, DBSCAN, Agglomerative clustering, and Gaussian Mixture Models (GMM), for clustering; Stage 2 uses geologically realistic synthetic Noddy models to examine both geological and mathematical collinearity, involving the N-Net trained on known lithology labels followed by SOM-based clustering; Stage 3 applies the methodology to field data with inherent collinearity patterns in Victoria, Australia, implementing additional spatial correlation analysis on SOM-based clustering.

2.1. Simple Synthetic Models - Isotropic Clusters

To establish a baseline understanding of SOM performance under controlled conditions of cluster separation and feature collinearity, we constructed a synthetic data set comprising two distinct isotropic clusters in bivariate feature space. Each cluster consisted of $n = 1000$ points randomly sampled from bivariate Gaussian distributions. Cluster 1 was centered at $(\mu_{pc11}, \mu_{pc21}) = (-d/2, -d/2)$ with equal standard deviations $\sigma_{pc11} = \sigma_{pc21} = \sigma$, while Cluster 2 consisted of n points drawn from a Gaussian distribution centered on (+d/2, +d/2), with standard deviations $\sigma_{pc12} = \sigma_{pc21} = \sigma$.

To introduce controlled collinearity, we derived a third feature, PC1b, from PC1 through the transformation PC1b = PC1 + ε, where ε represents Gaussian noise drawn from N(0, 0.02). This construction ensured strong correlation between PC1 and PC1b while introducing minor deviations. Additionally, we included two independent noise variables, Noise 4 and Noise 5, each drawn from standard normal distributions N(0, 1), to simulate irrelevant features commonly encountered in real-world data sets.

We systematically varied two parameters to control data set characteristics: the standard deviation σ, ranging from 0.2 to 0.8 to modulate intra-cluster dispersion, and the inter-cluster distance d, ranging from 0.5 to 6.5 to adjust cluster distinctiveness. Figure 1 illustrates the resulting feature relationships for representative values of μ and σ. The correlation matrix reveals the designed linear dependencies among input features, particularly the strong correlation between PC1 and PC1b, while Noise 4 and Noise 5 remain uncorrelated with the principal components.

2.1.1. UMLA Implementation

For simple synthetic models, we implemented a standard SOM algorithm with consistent parameters to facilitate direct comparison across our models. Our implementation used a 20 × 20 node architecture arranged in a rectangular grid, maintaining sufficient resolution to capture feature space complexity while avoiding excessive computational overhead. We conducted 1000 training iterations for each experiment, beginning with an initial learning rate of 0.3. A Gaussian neighborhood function was used to govern the adaptation of nodes surrounding the best-matching unit. Euclidean distance served as the key metric for winner determination. Weight vectors were initialized randomly within the range of input data values to ensure unbiased starting conditions. The Unified Distance Matrix (U-Matrix) was computed as the mean Euclidean distance between each SOM node's weight vector and its immediate neighbors, providing a visualization of cluster boundaries and topological structure within the trained SOM. The SOM implementation was based on the MiniSom library (Vettigl, 2025) with custom modifications to accommodate our experimental design requirements and output analysis procedures.

Further experimentation aimed at expanding the spectrum of ML to evaluate how other UMLAs perform with fixed configurations aligned with the two-cluster structure of our synthetic data sets. All expanded UMLAs utilized scikit-learn implementations (v1.2+) with parameters held constant across experiments (Pedregosa et al., 2011). We briefly describe the chosen algorithms and their relative implementation. K-Means is a partition-based clustering method (Milligan & Cooper, 1987). It groups data into k clusters by minimizing within-cluster sum of squared Euclidean distances to each cluster centroid (Ahmad & Dey, 2011; Jain, 2010). This corresponds to assuming clusters with equal isotropic covariance structure, so the method effectively partitions data into compact, isotropic groups. The methodology is sensitive to collinearity, as correlated features distort distance-based assignments (Albuslimi et al., 2021; Hajnajafi et al., 2021). As a centroid-based algorithm, K-Means requires pre-specification of cluster count (k). We set $k = 2$ to match the known generative structure of our synthetic data. The default initialization in scikit-learn was used to optimize centroid seeding, mitigating sensitivity to initial conditions. The algorithm minimizes within-cluster variance using Lloyd's iterative optimization (Portales et al., 2025), with convergence determined by a relative tolerance of 1e−4 and a maximum iteration of 300. DBSCAN is a density-based clustering method (Ester et al., 1996). It identifies clusters as high-density regions separated by low-density areas. It is robust to collinearity if the underlying density structure remains discernible but struggles with uniform density distributions (Elahifar & Hosseini, 2024; Özdemir et al., 2021). This density-based method requires no preset cluster count but is sensitive to distance scales. Input features were standardized via StandardScaler to ensure equal feature weighting. The key parameters, neighborhood radius and core point threshold, are set to 0.5 and 5, which were calibrated to separate high-density regions while minimizing noise in synthetic prototypes. The resulting labels were binarized using modulo 2 to align with our two-class evaluation

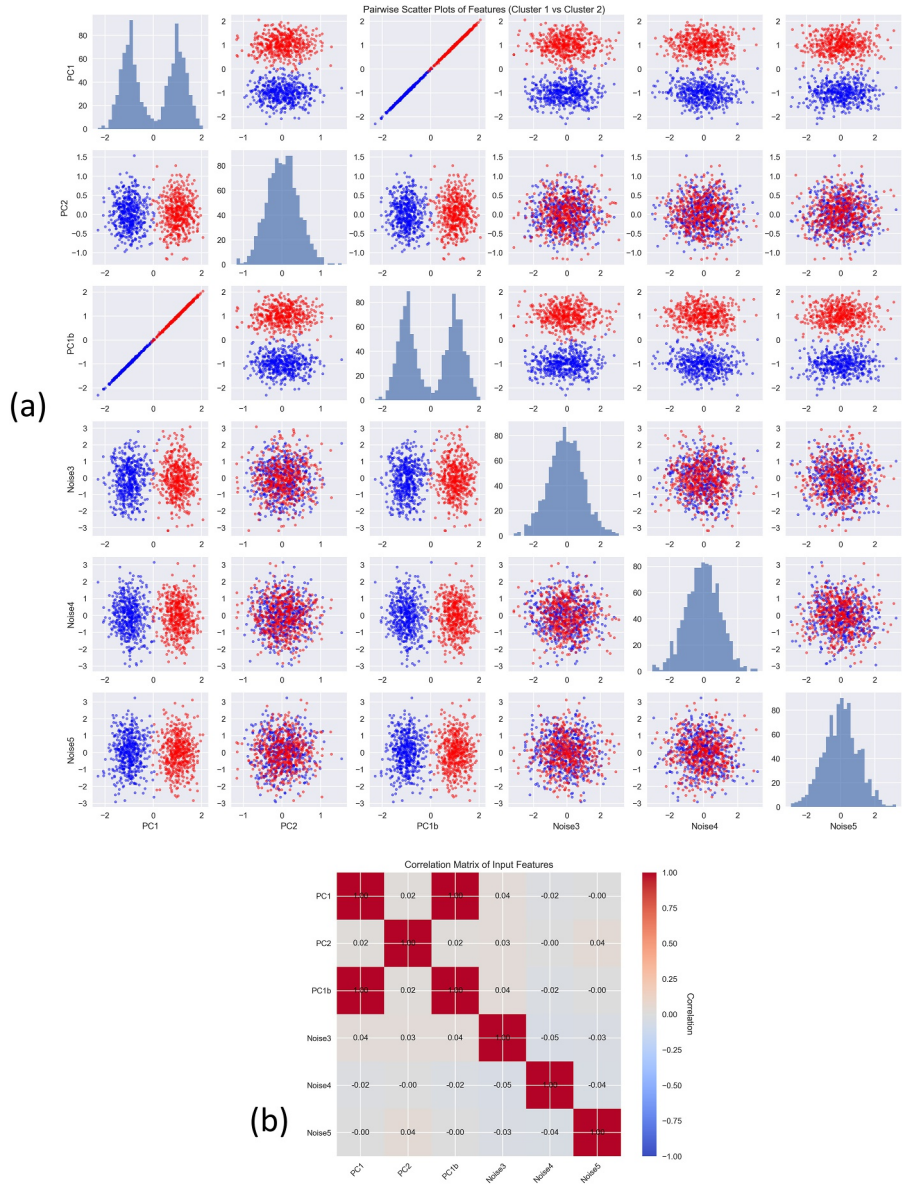

Figure 1. Visualization of feature relationships in Model 1 with cluster distance of 2.0 and standard deviation of 0.4. (a) Pairwise scatter plots of features colored by cluster membership. The diagonal presents frequency distributions for each feature by cluster membership, while off-diagonal panels display scatter plots of feature pairs colored by cluster assignment. (b) Correlation matrix of the input features. Note the ∼1.0 correlation between PC1 and PC1b demonstrating the intentional feature redundancy.

framework, converting noise points and cluster labels to binary outputs. Agglomerative Clustering is a hierarchical clustering method. It merges similar points iteratively, forming a dendrogram (Tokuda et al., 2022). While it captures nested structures, collinearity can distort linkage metrics, affecting cluster cohesion (Tokuda et al., 2022). Using a hierarchical approach, we fixed the number of clusters to 2 and applied Ward's linkage (Miyamoto et al., 2015) to minimize within-cluster variance during merging. Euclidean distance served as the affinity metric. No connectivity constraints or distance thresholds were applied. GMM is a probabilistic clustering method (Reynolds, 2009). It assumes that data is generated from a mix of Gaussian distributions. It accommodates elliptical clusters but may overfit with collinear features due to covariance matrix estimation issues. However, covariance matrix estimation can become unstable when feature collinearity induces near-singularity in the covariance matrix or when the ratio of samples to covariance parameters is unfavorable within individual mixture components (Boualla et al., 2015; Wallet & Hardisty, 2019). In our data sets, both factors contribute:

mathematical derivatives introduce strong feature correlations, while minor lithological units occupy limited spatial extents. The probabilistic framework assumed two Gaussian components. Full covariance matrices accommodated elliptical cluster shapes. Parameters were estimated via Expectation Maximization (Sammut & Webb, 2010). Convergence occurred when log-likelihood improvement fell below 1e−3 or after 100 iterations.

2.1.2. Performance Evaluation

Model performance was evaluated using two input configurations applied to the simple synthetic models: (a) Raw set: a simplified set containing only the uncorrelated primary features (PC1 + PC2), and (b) Extended set: a feature set incorporating collinear components and random noise (PC1 + PC2 + PC1b + Noise). Quantitative evaluation includes four standard metrics: overall accuracy (the proportion of correctly classified instances), precision and recall for each cluster, and the F1 score (the harmonic mean of precision and recall; De Diego et al., 2022; Gong, 2021). Since ground truth labels were known for the bivariate synthetic data, they served as a reference for comparison. To isolate the impact of collinearity, results were reported both individually and as in performance ratios between the raw and extended input sets. Quantitative results are reported from a single, representative run. The stability of these results was verified through a sensitivity analysis comprising 10 independent runs with different random initializations. Variation in core performance metrics is assessed using the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean and expressed as a percentage (Lovie, 2005).

2.2. Complex Synthetic Models - Noddy Geophysical Simulations

2.2.1. Synthetic Model Built-Up

We developed six distinct geological scenarios using Noddy (Jessell & Valenta, 1996) to generate synthetic geophysical responses, representing pseudo-realistic subsurface 3D geological models incorporating variable degrees of geological complexity.

Noddy is a 3D geological forward modeling package that builds synthetic models by sequentially applying geological events such as deposition, folding, faulting, and intrusion. It can also generate gravity and magnetic responses, making it a useful tool for testing inversion methods and developing synthetic geophysical data sets.

The six models, Q001–Q006, comprise 4–6 lithological units and 4–8 deformation events in their structural histories. The lithological units include igneous units such as mafic volcanic (basalt), felsic intrusive (granite), and intermediate intrusive (gabbro and diorite); regional metamorphic rocks including gneiss and schist; and sedimentary units comprising chert, carbonaceous rock, sandstone as psammitic sediment, mudstone as pelitic sediment, and unconsolidated sedimentary cover deposits. A set of structural brittle discontinuities, and ductile deformation was incorporated into the model, dominated by sine folds, plane faults, igneous intrusions, and tilting episodes.

We developed a Geological Complexity Index (GCI), which is normalized to a 0–1 scale, to quantitatively represent the model's complexity. The index follows a weighted linear aggregation approach commonly used in composite indicator construction (Greco et al., 2019; Nardo et al., 2008) and is conceptually consistent with the geodiversity framework (Lindsay et al., 2013), in which geological complexity is quantified through a weighted combination of multiple geological components to characterize model heterogeneity and structural variability.

$$\text{Index} = \sum_{i=1}^{n} w_i \frac{x_i}{M_i}$$

Where:

- x_i = observed value for component i,
- M_i = maximum (or normalization constant) for component i,
- w_i = weight of component i, with $\sum w_i = 1$.

For GCI:

- x = [LithologyCount, DeformationEvents, StructureTypes],

- M = [8,8,4],
- w = [0.5, 0.3, 0.2].

The weights were assigned based on the relative contribution of each component to computational complexity and geological heterogeneity, with the lithology count receiving the highest weight due to its direct influence on material property variations throughout the model domain. As shown in Table 1 and exemplified in Figure 2, the map of Model Q001 spanned a 1212 by 845 m area at a resolution of 1212 × 845 grid elements.

For each Noddy model, we computed two geophysical responses through forward modeling, resulting in (a) Total Magnetic Intensity (TMI) data, magnetic anomalies (from susceptibility contrasts and remanence), and (b) Gravity fields, gravity anomalies (from density contrasts). We then applied first-order vertical derivatives, analytical signals, and tilt transforms to the geophysical responses, which are commonly used in geophysical interpretation (Dentith & Mudge, 2014; Fairhead, 2015). The Tilt Angle, calculated as the arctangent of the ratio between the vertical derivative and the horizontal gradient magnitude, provided edge detection capabilities largely independent of amplitude. As shown in Figure 2, lithological units exhibit distinct but variable geophysical signatures, which directly influence SOM clustering results. Additional maps for scenarios Q001–Q006 are provided in the Figures S1–S6 of Supporting Information S1, and a working example of GCI is shown in Text S1 of Supporting Information S1.

2.2.2. Collinearity and PCA Analysis

We examined two collinearity types: (a) geologically driven collinearity, where multiple geophysical responses co-vary due to shared geological controls, and (b) statistically redundant collinearity from mathematical derivations, for example, tilt transforms of TMI data. For geologically driven collinearity, we calculated moving-window Pearson correlations (5 × 5 grid elements) between fundamental parameters (TMI vs. Gravity) and Variance Inflation Factors, using a VIF threshold of 5, and mapped these zones to local GCI from Noddy's ground truth. The local GCI is simplified to combine two equally weighted components: the Local Lithology Count and the Lithology boarder Length, where border length counts adjacent lithology changes within each window. Collinearity hotspots were identified when local-to-global correlation ratios exceeded the 90th percentile threshold, with results visualized as correlation heatmaps overlaid with GCI contours (0–1 scale) and statistically significant hotspots ($p < 0.05$).

Statistically redundant collinearity was assessed on the four geophysical parameters, including TMI, TMI Tilt, Gravity, and Gravity Tilt, for each lithological unit present in the models. The PCA serves two purposes: (a) visualization of feature relationships in reduced-dimension space, and (b) quantification of variance explained by principal components to assess the degree of collinearity within each geological setting. For each Noddy model, we generated PCA scatter plots projecting lithological units onto the Principal Component 1 (PC1) versus Principal Component 2 (PC2) space, where each rock type is represented by data points colored according to its lithological classification. To facilitate interpretation and reduce visual complexity, we also created simplified elliptical representations showing lithological unit centroids and their associated standard deviations as confidence ellipses.

2.2.3. SOM Implementation and Evaluation Framework

A fundamental challenge in applying SOMs to geological interpretation is the mismatch between UMLA outputs and geological reality. SOMs naturally partition data into a large number of distinct classes based on their similarities, yet geological models are constructed with a predetermined number of lithological units—ranging from 3 to 8 lithological units in our synthetic data sets. To address this challenge, we implemented two ML pipelines that work in concert: a feedforward neural network that acts as a geological knowledge enhancer and an SOM-based classifier that performs the final lithological mapping. Figure 3 provides a workflow diagram of this architecture.

The SOM implementation utilized a 25 × 25 grid architecture (625 nodes total), configured through systematic sensitivity analysis to capture the geophysical variations present in the synthetic data sets. Grid size selection involved evaluating configurations from 10 × 10 to 30 × 30 nodes, with classification accuracy and cluster separation serving as optimization criteria. Smaller grid configurations failed to adequately represent the subtle geological patterns observed in the data, while larger grids introduced computational overhead without

JGR: Machine Learning and Computation

10.1029/2025JH001107

Table 1
Geological Complexity Characteristics of Noddy Models

Model	Lithology count	Lithology types	Deformation events	Dominant structures	Key geological features	Purpose	GCI
Q001	4	Sedimentary cover, Psammitic sediment, Felsic intrusive, Pelitic sediment	3	Dyke intrusion, tilt, fault	Felsic intrusion in sedimentary sequence	Evaluate fault-induced collinearity	0.51
Q002	4	Metamorphic rocks, Sedimentary cover, Psammitic sediment, Pelitic sediment	4	Fault, tile, sine fold	Metamorphic-sedimentary units, planar fault	Test collinearity at simple lithological boundaries	0.55
Q003	4	Sedimentary cover, Psammitic sediment, Felsic intrusive, Pelitic sediment	4	Tilt, dyke intrusions, sine folds (2)	Multi-phrase folding with intrusion	Study superimposed deformation effects	0.55
Q004	7	Mafic intrusive, Sedimentary cover, Psammitic sediment, Felsic intrusive, Pelitic sediments	4	Sine folds (2), fault, tilt	Mafic-sedimentary interface, single fold-fault system	Test volcanic-sedimentary system collinearity	0.61
Q005	8	Intermediate intrusive, Chert, Sedimentary cover, Carbonaceous rock, Pelitic sediment, Psammitic sediment	4	Sine folds (2), tilt, gabbro intrusion	Intermediate intrusion with folded strata	Assess intrusion-related collinearity	0.68
Q006	9	Chert, Carbonaceous rock, Sedimentary cover, Psammitic sediment, Felsic intrusive, Pelitic sediment	8	Faults (2), tilts (2), sine folds (3), dyke intrusion	Chert-carbonate sequence, polyphase deformation	Examine extreme lithological/structural complexity	0.88

10.1029/2025JH001107

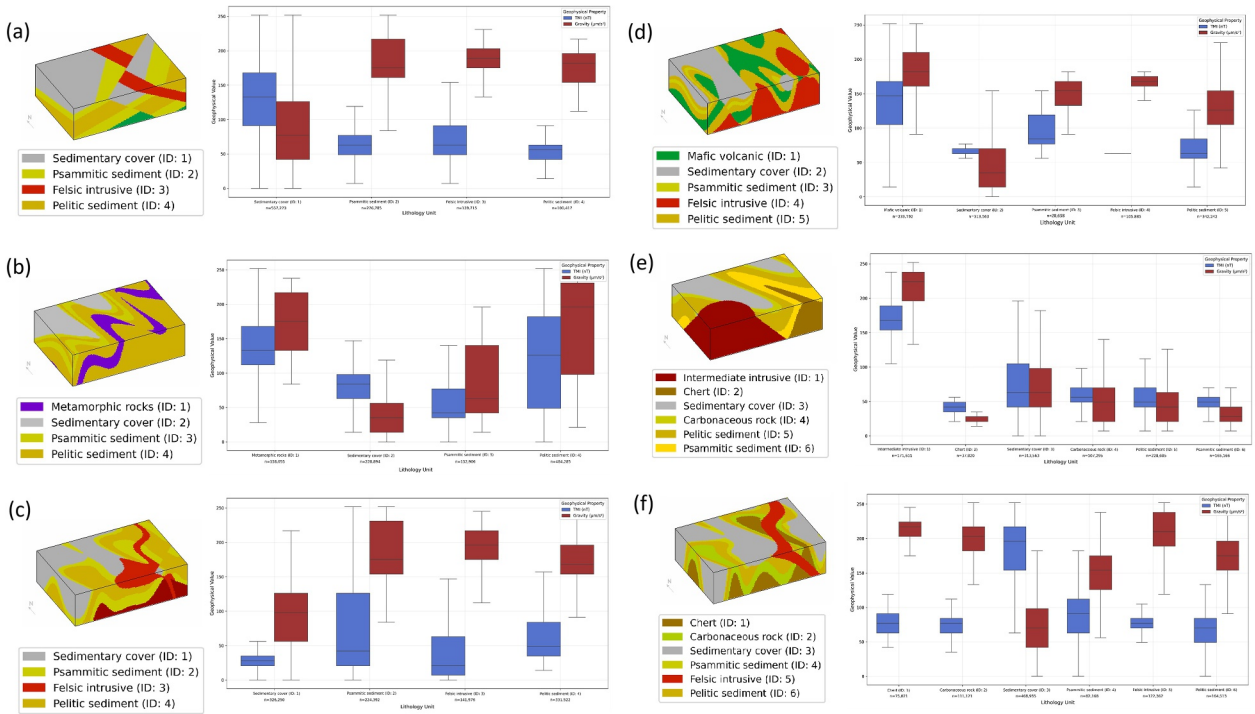

Figure 2. Visualization of the Noddy Model and their geophysical response quantile plot of the lithology units using percentiles (0, 10, 50, 90, 100). (a) Model Q001; (b) Model Q002; (c) Model Q003; (d) Model Q004; (e) Model Q005; (f) Model Q006. Each panel displays a 3D block model (1212 × 845 × 470 m) and corresponding geophysical response plots. TMI is shown in nanotesla (nT), and gravity response is in microgals (μm/s^2).

meaningful performance gains. The SOM training used 15,000 iterations with an initial learning rate of 0.5 and a sigma value of 1.5, applied to standardized input features. These hyperparameters were determined through systematic grid search across learning rates (0.1, 0.3, 0.5, 0.7) and sigma values (1.0, 1.5, 2.0, 2.5), with the selected configuration yielding a stable convergence.

The framework incorporates geological knowledge through a four-layer feedforward neural network (N-Net) that serves as a feature enhancement (Bebis & Georgiopoulos, 1994; Liang et al., 2023). This network was initially trained on the complete synthetic data set using an 8-dimensional geophysical feature vector, comprising raw magnetic and gravity measurements alongside their respective derivatives and transforms. Training used the Adam optimizer (Kingma & Ba, 2014) over 50 epochs with dropout regularization to prevent overfitting (Srivastava et al., 2014). The trained network generates probability distributions across lithological classes for each spatial location. These probability vectors are integrated with the original geophysical features, expanding the input space to 8 + n dimensions, where n represents the number of lithological classes for each geological model. This augmented feature set provides the SOM with both raw geophysical signatures and learned geological constraints, enabling unsupervised clustering that incorporates data-driven lithological patterns while preserving exploratory capabilities for identifying unknown geological structures. This approach renders the overall pipeline semi-supervised, as it combines supervised feature learning (N-Net training) with unsupervised spatial clustering (SOM).

Three progressively complex input configurations were evaluated to assess collinearity effects and feature complexity impacts. The first set used a minimal configuration with raw TMI and gravity data only, creating a 2-dimensional feature vector representing the most constrained input. The second set used an extended configuration combining TMI and gravity with their respective derivative transforms, producing an 8-dimensional feature vector that included raw TMI and gravity measurements, first vertical derivatives of both fields, tilt angle transforms of both fields, and analytical signal transforms of both fields. This configuration introduces mathematical collinearity through the derivative relationships between features. The third set applied PCA to the 8-dimensional feature set, retaining components explaining 95% of variance to address collinearity while preserving essential geophysical information.

10.1029/2025JH001107

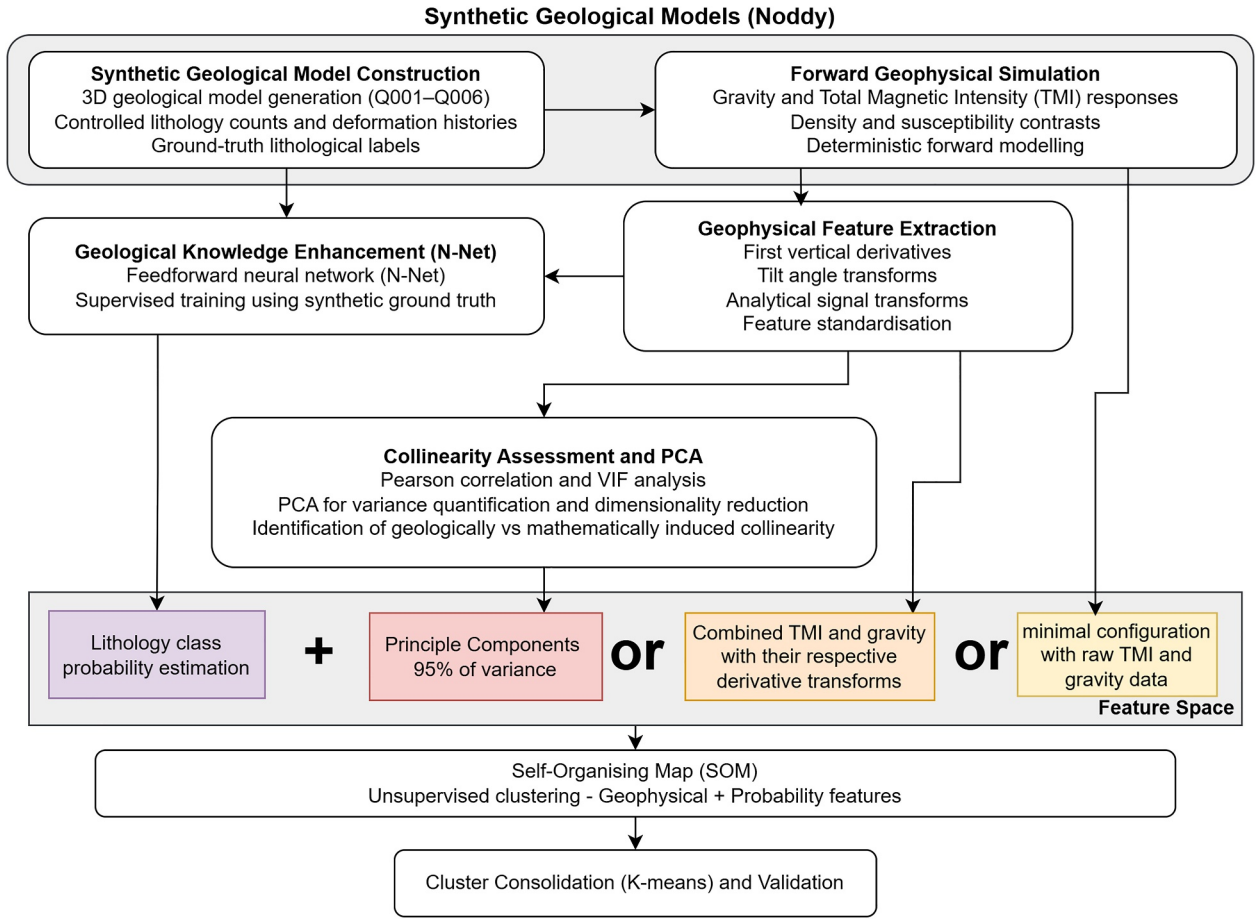

Figure 3. Visualization of the workflow.

The methodology employed a multi-stage process for lithological classification. Initial clustering begins with the enhanced SOM performing unsupervised clustering on the augmented feature space, identifying natural data groupings informed by both geophysical measurements and neural network-derived geological constraints. Following this initial clustering, K-means clustering aggregates SOM outputs to match the expected number of lithological units for each geological model, directly addressing the dimensional reduction challenge by consolidating the large number of SOM nodes into geologically meaningful lithological classes. To balance the deterministic clarity and automation provided by K-means against the risk of partially disregarding the SOM's topological structure, the final stage involves lithological assignment using maximum likelihood estimation, where each cluster adopts the lithology class most frequently represented in the corresponding ground-truth spatial region.

Classification performance was evaluated against synthetic ground-truth lithological maps from Noddy geological simulations. Primary performance indicators included overall classification accuracy from pixel-wise comparison between predicted and ground-truth maps, Adjusted Rand Index measuring agreement between cluster assignments and true lithological boundaries (Hubert & Arabie, 1985), and Silhouette coefficient assessing cluster separation quality (Rousseeuw, 1987). Detailed performance analysis incorporated confusion matrices providing class-wise performance assessment with both raw counts and normalized proportions, F1-scores for individual lithological classes that balance precision and recall, addressing class imbalance, and confidence measures quantifying the proportion of pixels within each cluster belonging to the assigned dominant lithology. Spatial analysis components examined misclassification mapping to identify systematic spatial error patterns and per-class precision and recall metrics to evaluate performance variability across lithological units.

2.3. Real-World Case Study

To contextualize our synthetic model results, we present a comparative analysis of geophysical signatures from real geological units in eastern Victoria, Australia. Following Xu et al. (2025b), this region provides an established reference data set with known lithological classifications. The SOM was implemented using a 10×10 hexagonal grid, trained for 200 epochs until convergence, with Euclidean distance as the similarity metric. No new clustering or lithological mapping was performed; rather, we analyze the geophysical feature relationships within existing geological classifications to compare signal characteristics between real and synthetic rocks.

The geophysical survey incorporated eight data sets at 50×50 m resolution: TMI (Geological Survey of Victoria, 2008), magnetic first vertical derivative (1VD), magnetic analytic signal (AS), potassium, thorium, total gamma-ray count, uranium (Minty et al., 2009; Skladzien & Bibby, 2005), and gravity measurements (Wynne & Bacchin, 2009). The lithological classification encompasses 111,319 rock samples collected across a 330×115 km region (Willman et al., 2005), representing 12 distinct geological units (Xu et al., 2025c), including psammitic and pelitic sedimentary rocks, chert (quartz-rich sedimentary units), felsic and intermediate/mafic/ultramafic (i/m/u) intrusive and volcanic rocks, as well as contact and regional metamorphic units. This multiparameter data set is standardized to reflect realistic patterns of collinearity driven using geological processes and interrelated measurement responses. Additional maps for the Victoria case study are provided in the Figures S6 of Supporting Information S1.

We implemented a dual-scale analytical approach for comparison with our synthetic models (Section 2.2). The regional analysis encompassed the full 330×115 km zone, comprising 6600×2300 grid elements at 50 m resolution. The local analysis involved one 60×40 km focus area (1200×800 grid elements at 50 m resolution) selected for geological similarity to Noddy models. These local zones capture distinct lithological configurations mirroring Q001–Q006 scenarios while maintaining consistent analytical resolution. We implemented R^2, the coefficient of determination (Dangeti, 2017), to examine the relationship between PCA-derived feature distances and geographic distances for each lithological unit, as it quantifies the proportion of variance in geographic distances explained by PCA-derived feature distances implemented by spatial correlation analysis. This analysis quantifies whether geophysical similarity measured in PCA space correlates with spatial proximity, providing insights into the spatial coherence of lithological classifications. For each pair of grid points within the same lithological unit, we calculated: (a) Euclidean distance in PCA space using the first two principal components, and (b) geographic distance based on easting and northing coordinates. High coefficient values indicate strong spatial coherence, where geophysically similar locations are geographically clustered, while low correlation suggests heterogeneous geophysical responses within lithological units.

3. Results

3.1. Collinearity Robustness in UMLA: Literature Synthesis

The SOM's advantageous properties are mathematically verifiable, and a worked example is shown in Text S2 of Supporting Information S1. While traditional techniques such as linear regression, the solution requires computing $\boldsymbol{w} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$, where singularity issues arise when columns of the design matrix X are linearly dependent. SOMs instead use competitive learning based on Euclidean distance. The winning node i^* for input $\boldsymbol{x}$ is determined by $i^* = \underset{i}{\operatorname{argmin}}\|\boldsymbol{x} - \boldsymbol{w}_i\|$, where $\boldsymbol{w}_i$ represents weight vectors. This L_2-norm minimization remains well-defined regardless of the rank of $\boldsymbol{x}$, making SOMs inherently robust to multicollinearity (Yin, 2008).

The neighborhood function's mathematical formulation explains SOMs' resilience to noise and redundancy. The weight update rule $\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \alpha(t)\,h_{i*i}(t)\,(\boldsymbol{x} - \boldsymbol{w}_i(t))$ incorporates a kernel function $h_{i*i}(t)$, for example, Gaussian, that smoothly distributes updates across neighboring nodes. Let input $x = s + n$, where s is the shared correlated signal and n is noise. For two correlated inputs $x_1 = s + n_1$ and $x_2 = s + n_2$, the update becomes $\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i + \alpha(t)\,h_{i*i}(t)\,(2s + n_1 + n_2 - 2w_i)$. Through this formulation, we observe that signal components s are reinforced while noise terms $n_1 + n_2$ tend to cancel to zero mean. This is akin to an implicit variance reduction mechanism.

From an optimization perspective, SOMs minimize a topological distortion measure (Kohonen, 2001): $\sum_t \sum_i h_{i*i}(t)\|\boldsymbol{x} - \boldsymbol{w}_i\|^2$. This is fundamentally different from PCA's objective of minimizing reconstruction error

$\left\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}\right\|^2$. By the Eckart-Young theorem (Golub et al., 1987), the Frobenius norm error $\left\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}\right\|_F^2$ equals the sum of squared discarded singular values $\sum_{i=k+1}^{n}\sigma_i^2$. Here, $\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}$ is the rank k approximation of $\boldsymbol{X}$. When collinearity exists, some $\sigma_i \approx 0$, the reconstruction error becomes dominated by these near-zero singular values and the pseudoinverse operations in PCA involving $\frac{1}{\sigma_i}$ amplify noise, making the solution numerically unstable. In contrast, SOMs are agnostic because their optimization operates directly on input-space distances without requiring full-rank covariance matrices (Flexer, 1999).

Empirical validations align with these mathematical insights. Yin (2008) showed that the SOM algorithm's Lyapunov stability, which could converge to a topology-preserving mapping even when rank $(\boldsymbol{x}) <$ dim $(\boldsymbol{x})$. Flexer's (1999) clustering experiments on rank-deficient data sets further confirmed that SOMs maintain approximately constant quantization error $Q = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x} - \boldsymbol{w}_i\|^2$, where N is the number of input $\boldsymbol{x}$ regardless of the degree of collinearity. Some empirical work also suggests that SOMs may adaptively down-weight redundant inputs during training (Kiviluoto, 1998). For PCA, the reconstruction error grows as $\frac{1}{\sigma_k}$ for small singular values σ_k of the data matrix $\boldsymbol{x}$. However, in most applications, PCA or factor analysis is applied beforehand to explicitly address multicollinearity (Nourani et al., 2013).

The ability to handle collinear features is not unique to SOMs but extends to other UMLAs. Collinearity robustness is not algorithm-specific, but rather arises from computational primitives, including distance comparisons, local updates, and topological objectives (Hastie et al., 2001; Jain et al., 1999):

1. Distance-Based Methods Avoid Matrix Inversion: Clustering algorithms that rely solely on pairwise distances, for example, Euclidean and Manhattan, or topological relationships, are invariant to linear dependencies in features because they operate on the input space rather than a parameter space requiring matrix inversion. For example, in the Agglomerative Clustering method, the linkage criterion, single, complete, Ward, etc., computes pairwise distances between points or clusters, for example, $d(\boldsymbol{x}_i,\boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$ (Ramos Emmendorfer & de Paula Canuto, 2021). Collinear features do not alter the relative distances as they contribute redundantly to the norm. For example, if $\boldsymbol{x}_i = [v_i, v_i]$ and $\boldsymbol{x}_j = [v_j, v_j]$ where the input shows perfect collinearity, the distance reduces to $\sqrt{2}|v_i - v_j|$, preserving ordinal relationships. Similarly, as previously shown, the winner selection in the SOM $\underset{i}{\operatorname{argmin}}\|\boldsymbol{x} - \boldsymbol{w}_i\|$ depends only on the input-space geometry, not the rank of the design matrix. In contrast, GMMs rely on covariance matrices $(\boldsymbol{X}^T\boldsymbol{X})$ or eigen decompositions, which fail when rank $(\boldsymbol{X}) <$ dim $(\boldsymbol{X})$
2. Noise Averaging in Local Updates: Algorithms that distribute learning across local neighborhoods, such as SOMs and hierarchical clustering, implicitly average out noise in collinear features while reinforcing consistent signals. For example, in hierarchical clustering, Ward's method merges clusters by minimizing variance within them (Murtagh & Legendre, 2014). For collinear features, Variance $= \sum_{i=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2$, where $\boldsymbol{\mu}$ is the cluster mean. Redundant features contribute identical directional errors, effectively weighting the true signal more heavily. Similarly, in SOM, the neighborhood kernel $h_{i*i}(t)$ smooths updates, as previously derived.
3. Topology Preservation: In clustering, topology refers to the preservation of neighborhood relationships, wherein data points that are proximate in the original feature space remain spatially close within the clustered representation. This concept is analogous to maintaining the relative spatial configuration of geological formations during dimensionality reduction, such as when projecting a three-dimensional structure onto a two-dimensional plane. Clustering methods that prioritize topology, for example, SOMs, hierarchical clustering, and DBSCAN, over linear separability are inherently robust to collinearity because they preserve ordinal or structural relationships rather than optimizing linear projections. In hierarchical clustering, the dendrogram structure relies solely on relative distances, unaffected by whether these distances result from independent or collinear features. The cophenetic correlation—a measure of how well the dendrogram preserves input distances—remains invariant under linear transformations of collinear features (Liu et al., 2022; Smith & Dubes, 1980). Comparably, the topographic mapping error $h_{i*i}(t)\|\boldsymbol{x} - \boldsymbol{w}_i\|^2$ in SOM penalizes neighborhood violations with non-linear dependencies.

Clustering algorithms differ in how they handle pairwise distances or density estimates, both of which can be affected by redundant or highly correlated features. The degree of this influence varies across methods. In terms of empirical implications, the agglomerative clustering methods although theoretically based on pairwise distances

that ignore variable dependencies, are nonetheless affected by feature collinearity. Studies have shown that cophenetic correlation coefficients and silhouette scores often improve when correlated features are removed or transformed (Hastie et al., 2009; Meigoony et al., 2014; Shahrestani & Sanislav, 2025). For example, studies consistently show that applying PCA prior to K-means clustering improves clustering validity indices and helps recover the underlying cluster structure in the presence of multicollinearity (Ghezelbash et al., 2023; Jain, 2010; Jooshaki et al., 2021). Consequently, hierarchical clustering is not inherently robust to collinearity, and pre-processing steps to reduce feature redundancy are commonly recommended. DBSCAN is comparatively less sensitive to collinearity. Some reports caution that duplicated or highly correlated dimensions can complicate the selection of density thresholds (ϵ) and skew cluster boundaries (Ester et al., 1996; Parsons et al., 2004; Saxena et al., 2017). While DBSCAN demonstrates reasonable empirical robustness, its reliability improves when redundant dimensions are removed or compressed (Li et al., 2023; Perafan-Lopez et al., 2022). Similarly, empirical evidence suggests that full-covariance GMMs perform well when such correlations are present (Murphy, 2012). However, when data dimensionality is high or the number of observations is limited, dimensionality reduction techniques or regularization are typically used to stabilize estimation (Facchinelli et al., 2001; Wilks, 2019).

3.2. Simple Synthetic Models: Baseline Collinearity Effects

The result from a single representative run is presented in Figure 4. It revealed systematic patterns in the degradation of SOM performance attributable to the distinctiveness of the clusters. Figure 4 illustrates a triphasic relationship between cluster separation and collinearity effects. The most pronounced performance differential at minimal separation (distance 0.0–0.5), the extended feature set outperformed the raw set, suggesting that collinear features may provide a compensatory signal when clusters overlap. This beneficial effect disappeared at moderate distances (1.0–2.0), where performance ratios declined slightly to 0.95–0.99, indicating a decrease in initial sensitivity to feature redundancy. Beyond this range (distance >2.0), the ratios stabilized at 0.99–1.0, as sufficient separation rendered collinearity irrelevant. These patterns are visually corroborated by UMAP projections (Figures 3a and 3b), where collinear features introduce measurable boundary uncertainty at intermediate distances while maintaining overall topology. The exact performance of the SOM is shown in Figure S8 of Supporting Information S1.

Similarly, Figure 5 shows how K-Means, DBSCAN, Agglomerative clustering, and Gaussian Mixture Models (GMM) respond to collinearity challenges. Unlike SOMs, these algorithms exhibited less consistent performance patterns. K-Means and GMM showed mild sensitivity to collinear inputs, with small gains at minimal separation but degradation at intermediate cluster standard deviation. DBSCAN and Agglomerative Clustering were more robust, showing degradation at low distances and high standard deviation.

3.3. Complex Synthetic Models: Geological Structure Influence

Figure 6a illustrates the comparison of classification results in Noddy Model Q002 in three feature configurations: Raw, All, and PCA-processed. The results suggest that the derived features, despite collinearity, can be beneficial when paired with appropriate preprocessing, reinforcing their utility in geological interpretation. Raw data provided reasonable but incomplete boundary delineation, while derivatives improved edge definition at the cost of noise in homogeneous regions. PCA-processed features delivered the most balanced results, maintaining sharp boundaries while minimizing misclassification in uniform lithologies. Spatial correlation mapping (Figure 6b) identifies two dominant collinearity regimes: (a) geologically driven collinearity ($\rho > 0.6$) concentrated at structural boundaries (e.g., fold zones in Q002 with $GCI_{local} > 0.8\ GCI_{global}$), and (b) statistically redundant collinearity prevalent in homogeneous regions, particularly for tilt-derived features. Classification results demonstrate that both pelitic sedimentary and sedimentary cover inference benefits from collinearity in low-GCI zones, while PCA processing optimally mitigates redundant collinearity in pelitic sedimentary and intrusive inferences in high-GCI regions, reducing uniform lithology misclassification. Additional maps for scenarios Q001 to Q006 are provided in the Figures S8–S12 of Supporting Information S1.

Across the 6 Noddy models, feature space organization (Figure 7) shows that geological complexity directly impacts cluster separability—low GCI_{global} exhibit well-separated lithology ellipses (<40% overlap), whereas high GCI_{global} show >80% intersection between units such as pelitic sediments and carbonaceous rocks, for example Figure 7f. Igneous units show tight, well-defined ellipses with minimal overlap, while complex

10.1029/2025JH001107

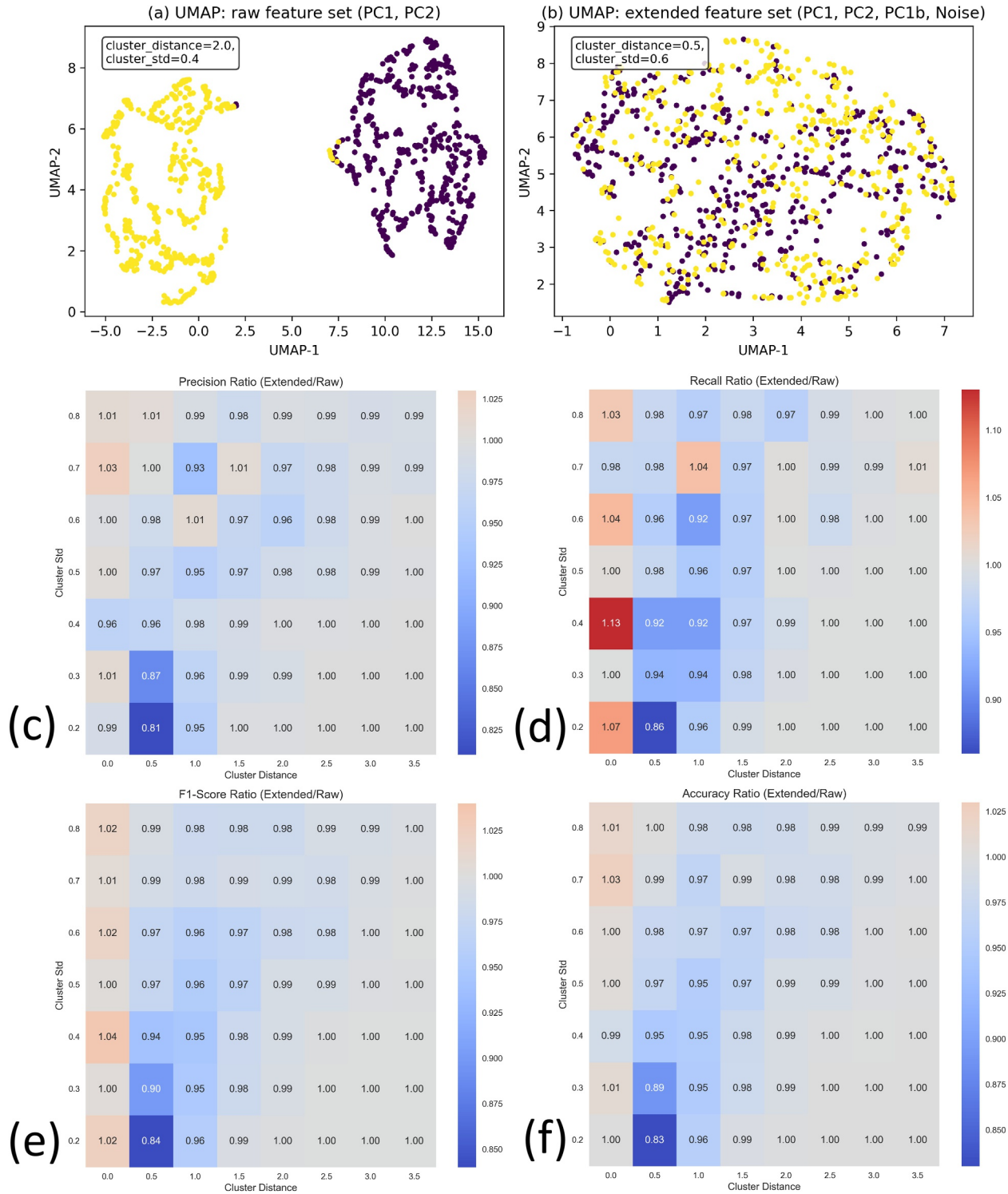

Figure 4. SOM classification results for the simple synthetic model. (a) Two-dimensional Uniform Manifold Approximation and Projection (UMAP) showing the geometric distribution of the raw feature set (PC1 + PC2) with cluster distance = 2.0 and standard deviation = 0.4. UMAP is a manifold learning technique used to visualize high-dimensional structure while preserving local and global relationships; (b) UMAP projection of the extended feature set (PC1 + PC2 + PC1b + Noise) with cluster distance = 0.5 and standard deviation = 0.6; (c) performance ratio comparison across cluster parameters, precision ratio, (d) recall ratio, (e) F1-score ratio, (f) accuracy ratio. In panels (c–f), the *x*-axis shows cluster distance (0.5–3.5), the *y*-axis shows standard deviation (0.2–0.8), and the color map indicates metric performance ratio, with values <1.0 indicate performance degradation with collinear features, values ≈1.0 show neutral effects, and values >1.0 demonstrate beneficial collinearity effects.

10.1029/2025JH001107

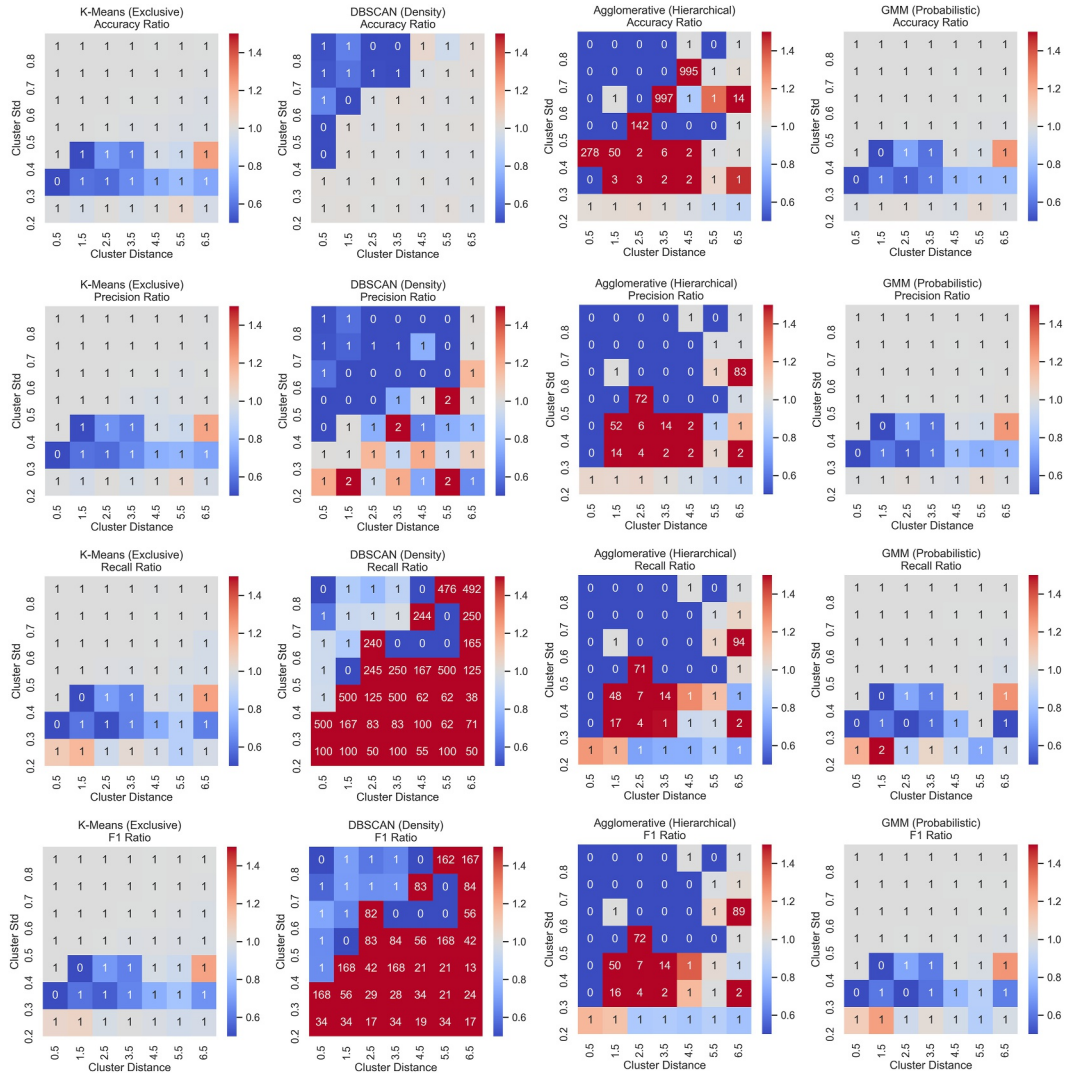

Figure 5. Comparison of UMLS's Robustness to Collinearity. Rows (top to bottom): Accuracy, Precision, Recall, F1-Score; Columns (left to right): K-Means, DBSCAN, Agglomerative clustering, GMM. All panels share the same spatial scale, with cluster distance (x) ranging from 0.5 to 6.5, cluster std (y) ranging from 0.2 to 0.8, and the color map indicating metric performance (0.6–1.4 scale) and the actual performance labeled on the grid.

structural models exhibit larger ellipses with significant overlap, indicating greater geophysical similarity across different lithological units. Table 2 reveals a systematic relationship between geological complexity (quantified by GCI) and the effects of collinearity on lithological classification. Three distinct regimes emerge from the data: High-complexity models (Q006-GCI = 0.88, Q005-GCI = 0.68) exhibit extreme collinearity in fundamental parameters (VIF values up to 19 for TMI in Q005), with PC1 accounting for 91%–96% of the variance. However, these models achieve relatively modest overall accuracy (Q006: 48%–49%, Q005: 54%–59%), with PCA processing providing consistent but incremental improvements. The high VIF values represent meaningful geophysical covariance, as evidenced by PCA's ability to enhance performance for specific lithologies, particularly intermediate intrusive in Q005 (from 0% to 74% F1-score). Intermediate-complexity models (GCI 0.55–0.61) demonstrate more substantial benefits from preprocessing strategies. Q004 shows notable improvement with PCA processing (from 61% to 66% overall accuracy), with mafic volcanics gaining significantly (from 60% to 71% F1-score). Interestingly, only the Mafic volcanic unit appears to benefit from PCA processing, as the F1-score for the felsic intrusive unit dropped dramatically—from 62% with all inputs to 0% when using PCA-processed inputs. Q002 and Q003 exhibit progressive accuracy gains moving from raw data through derivative inclusion to PCA processing (Q002: 57% → 58% → 59%; Q003: 45% → 48% → 52%). The low-complexity

10.1029/2025JH001107

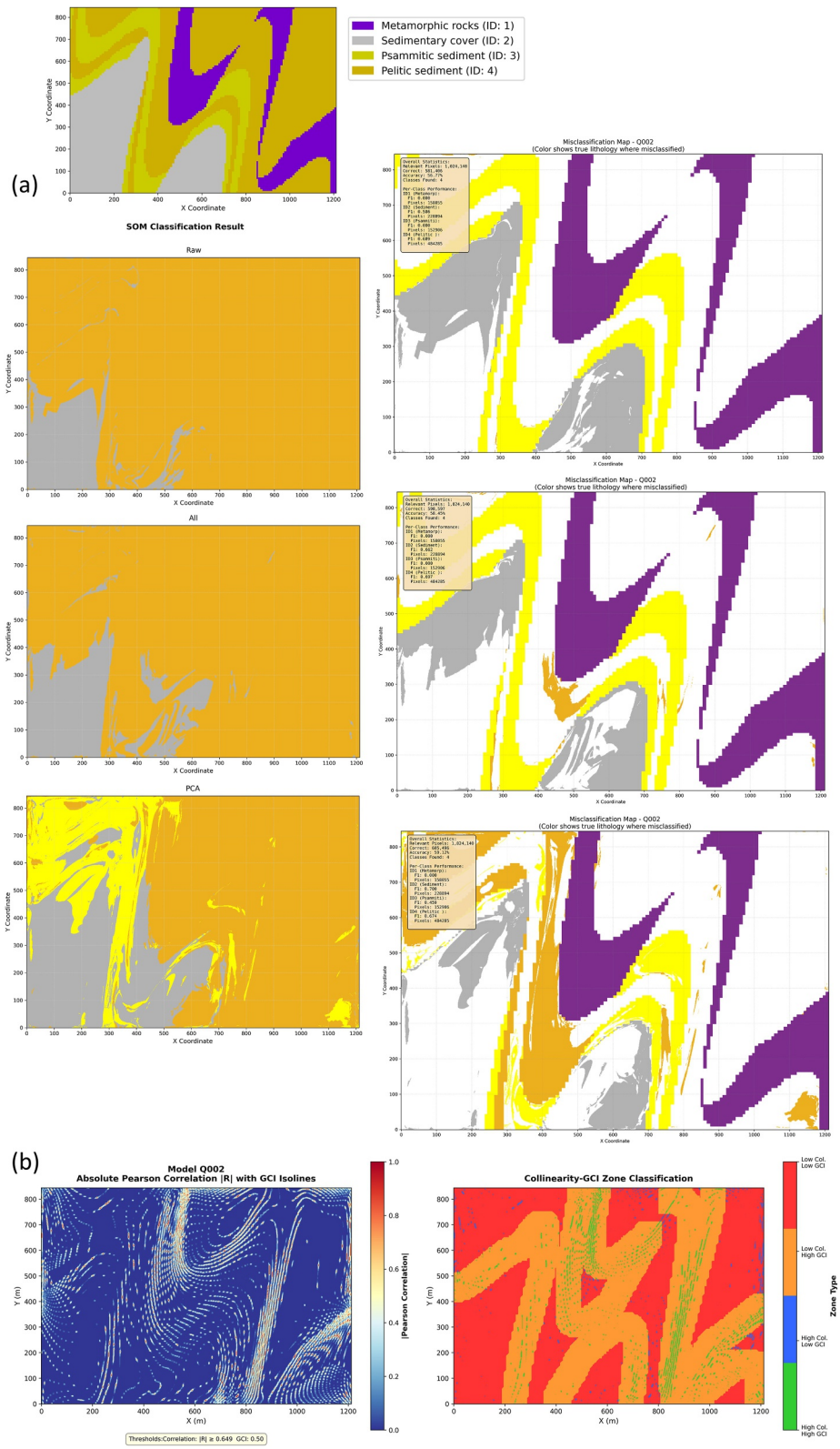

Figure 6.

model (Q001-GCI = 0.51) exhibits the highest overall performance, achieving 70% accuracy with PCA processing compared to 61% with raw data. This model shows minimal collinearity (VIF < 2 throughout) and demonstrates that preprocessing becomes increasingly beneficial as feature complexity increases. Lithological responses vary markedly by unit type and geological context. Sedimentary cover units generally maintain consistent performance across all models (59%–84% F1-scores), while specialized units show model-specific sensitivity. Pelitic sediments respond favorably to PCA processing in most cases, mafic volcanics benefit substantially from enhanced feature sets in Q004, and intermediate intrusive shows dramatic improvement with derivative processing in Q005. Chart and metamorphic rocks consistently achieve zero F1-scores across all input configurations in multiple models, indicating fundamental classification challenges that transcend preprocessing strategies. These variations show that processing effectiveness and classification performance depend on both geological complexity and lithological assemblage. Notably, the results reveal a novel approach to map collinearity effects across distinct geological domains, providing insights into the interplay between geological structure and geophysical response.

3.4. Real-World Analysis: Local Collinearity Variations

Real-world lithological clustering shows significantly greater overlap compared to synthetic models, with several rock types occupying similar regions in the PCA space. As shown in Figure 8a, Principal component analysis reveals PC1 (51.8%) and PC2 (19.2%) as the dominant variance components, with subsequent components PC3-PC5 showing progressively smaller contributions with <10% variance each (full breakdown in Figures S13 of Supporting Information S1). Sedimentary units (psammitic and pelitic rocks) form partially overlapping clusters, indicating similar geophysical signatures despite different geological origins. Intrusive units show moderate separation, with felsic and i/m/u rocks forming distinct but adjacent clusters. Metamorphic units closely resemble contact metamorphic rocks, and both occupy a central position in feature space with extensive overlap with other lithologies. Figure 8b illustrates that real-world lithological units have larger standard deviations compared to synthetic models, indicating greater geophysical heterogeneity within individual rock types. This heterogeneity reflects the natural complexity of geological processes, including variable mineral compositions, alteration effects, and structural modifications that are not fully captured in synthetic models. Spatial patterns in the study area in Figure 8c align with the PCA-derived clusters. The documented spatial variability in rock sample density (Xu et al., 2025b) shows that sedimentary rocks, particularly pelitic and psammitic types, exhibit a geometric clustering pattern, whereas igneous and metamorphic units display more irregular spacing.

Local collinearity variations across geography in Figure 7d mirror the heterogeneity trends observed in Figure 7b. Principal component analysis reveals PC1 (39.4%) and PC2 (23.2%), with subsequent components PC3-PC5 showing progressively smaller contributions with <20% variance each with the full breakdown in Figure S14 of Supporting Information S1. Nevertheless, the summarized R^2 for each of the 12 lithological units in both eastern Victoria and the focused area indicate weak to no relationships between PCA-derived feature distances and geographic distances for most lithological units, as shown in the Figures S15 and S16 of Supporting Information S1.

4. Discussion

4.1. Geological Context Determines Collinearity Impact

Our results demonstrate that collinearity effects are governed by cluster separability, which aligns with theoretical expectations.

In the isotropic cluster experiments, collinearity effects were governed by the degree of separation between clusters. When clusters were minimally separated (distance <0.5), collinear features provided compensatory

Figure 6. Lithological classification and collinearity analysis for Noddy Model Q002, $GCI_{global} = 0.55$. (a) Classification results showing (top to bottom): Ground truth geology; raw input clusters; all input clusters; PCA-processed input clusters with component plots and predicted lithology maps below each column. Displayed from left to right are the predicted lithology maps and the misclassification map. In the latter, white regions denote grid elements that were correctly classified. (b) Spatial Collinearity Analysis: Enhanced correlation analysis between TMI and gravity responses presented as isolines to emphasize spatial continuity of geophysical relationships. Absolute Pearson correlation coefficients (|R|, color scale 0–1) were used to represent true collinearity regardless of positive or negative correlation direction, as both indicate statistical dependence. All maps share the same 100 m increment spatial scale (easting x, northing y).

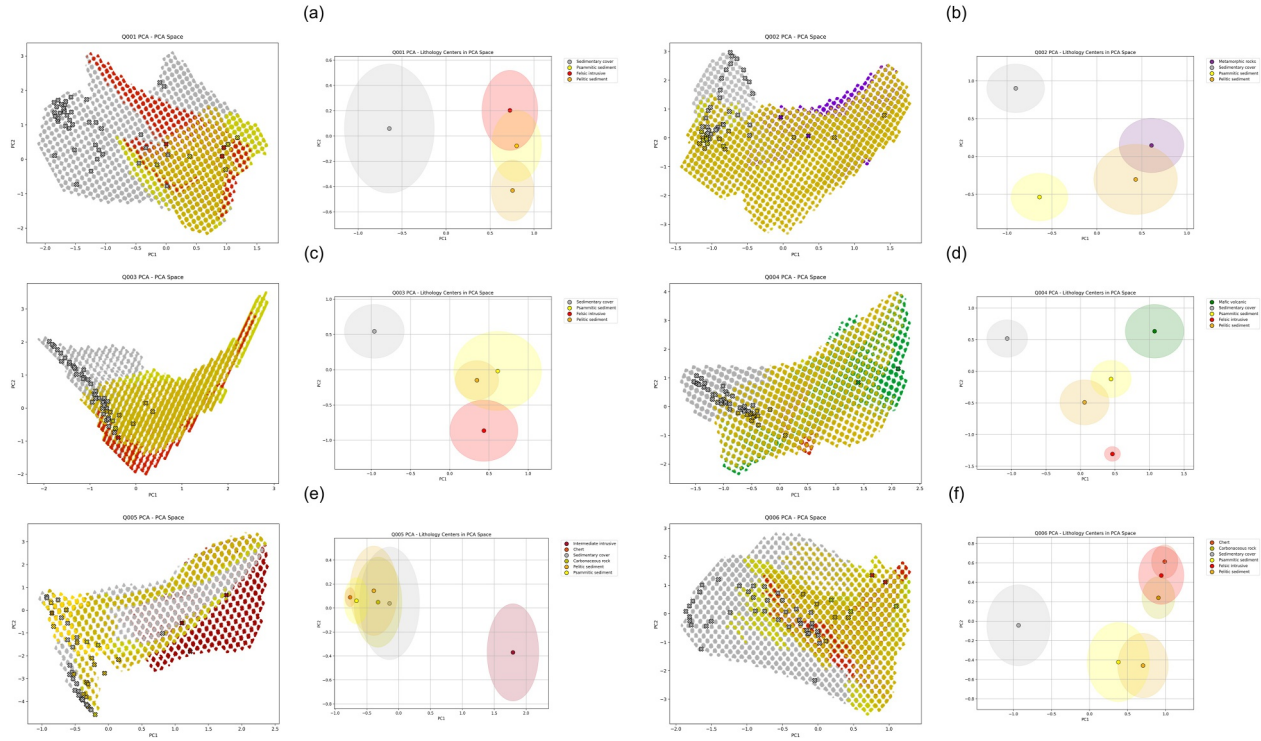

Figure 7. PCA representations of six Noddy synthetic models (Q001–Q006; panels a–f). For each model, the left panel displays a scatter plot of lithology types in PCA space (PC1 vs. PC2), while the right panel presents a simplified visualization with ellipses indicating the mean and standard deviation of each lithology cluster. The X markers indicate the centers of the SOM nodes derived from the clustering experiment using PCA-processed inputs. Axis labels correspond to principal components (PC1: *x*-axis; PC2: *y*-axis), with colors assigned according to the legend.

signals that improved classification accuracy, suggesting that redundancy can mitigate ambiguity in overlapping distributions. This benefit diminished as inter-cluster distance increased, with performance ratios stabilizing near unity for well-separated clusters. Notably, topology-preserving algorithms like SOMs and distance-based methods (DBSCAN, Agglomerative Clustering) exhibited inherent robustness to collinearity, outperforming parametric methods (K-Means, GMM) that rely on covariance structures.

The Noddy simulations revealed that collinearity effects are modulated by geological complexity (quantified via GCI). In low-GCI models, mathematically redundant collinearity, for example, derivative-induced correlations, could be safely reduced through PCA without information loss. In contrast, high-GCI models demonstrated that geologically driven collinearity reflecting shared geological discontinuities (either structural, lithological or petrophysical) enhanced classification when preserved, particularly at lithological boundaries. Intermediate-GCI regimes required a balanced approach where selective PCA improved discrimination of sedimentary units while the benefits of expanded feature sets became apparent for igneous classifications.

The Victoria case study reveals how geological complexity manifests as both spatial partitioning and organizational heterogeneity within lithological domains. Geological complexity, in this context, extends beyond the number of lithological units or structural features—it represents the degree of partitioning and the internal organization of geophysical properties within and across geological boundaries. The real-world data exhibited pervasive within-unit geophysical heterogeneity, evidenced by weak spatial coherence ($R^2 < 0.1$) between PCA-derived and geographic distances, contrasting sharply with the relatively homogeneous synthetic lithological units.

These observations align with the multi-step iterative clustering approach described by Xu et al. (2025b), where highly processed data establish broad geological domains based on regional trends, moderately processed data refine lithological indicators, and simply processed data capture fundamental geophysical properties. Their strategy of excluding "clean" cluster nodes at each step and progressively refining mixed components mirrors our

Table 2
Performance Metrics for Different Input Configurations Across 6 Noddy Models

Model #	Input	Overall accuracy (%)	Adjusted rand score	Silhouette score	Sedimentary cover (%)	Psammitic sediment (%)	Pelitic sediment (%)	Metamorphic rocks	Chert	Carbonaceous rock	Felsic intrusive	Intermediate intrusive	Mafic volcanic
					F1-Score								
Part (a) Classification scoring													
Q001	Raw	61	0.31	0.14	84	63	0	–	–	–	34%	–	–
Q001	All	66	0.38	0.15	84	56	50	–	–	–	0%	–	–
Q001	PCA	70	0.42	0.12	84	63	0	–	–	–	34%	–	–
Q002	Raw	57	0.10	0.04	59	0	69	0%	–	–	–	–	–
Q002	All	58	0.14	0.04	66	0	70	0%	–	–	–	–	–
Q002	PCA	59	0.24	0.08	70	46	67	0%	–	–	–	–	–
Q003	Raw	45	0.04	0.24	56	0	54	–	–	–	0%	–	–
Q003	All	48	0.16	0.12	66	0	49	–	–	–	0%	–	–
Q003	PCA	52	0.19	0.08	69	45	48	–	–	–	0%	–	–
Q004	Raw	61	0.36	0.19	81	0	50	–	–	–	0%	–	60%
Q004	All	62	0.28	0.04	74	0	53	–	–	–	62%	–	56%
Q004	PCA	66	0.36	0.18	77	0	61	–	–	–	0%	–	71%
Q005	Raw	54	0.28	0.15	63	45	54	–	–	31%	–	0%	-
Q005	All	56	0.32	0.15	69	44	48	–	0%	0%	–	71%	–
Q005	PCA	59	0.31	0.13	69	51	56	–	0%	0%	–	74%	–
Q006	Raw	48	0.13	0.05	67	0	0	–	0%	40%	13%	–	–
Q006	All	48	0.03	0.02	64	0	0	–	0%	0%	33%	–	–
Q006	PCA	49	0.05	0.09	65	0	4	–	0%	0%	40%	–	–

Model number	GCI_{global}	Mean GCI_{local}	VIF TMI	VIF gravity	VIF TMI derivative	VIF gravity derivative	PC1 variance explained (%)	PC2 variance explained (%)
Part (b) Complexity indexes								
Q001	0.51	0.065	1.8	1.8	1.43	1.39	80.90	19.10
Q002	0.55	0.067	8.8	8.4	1.29	1.43	88.80	11.20
Q003	0.55	0.065	3.4	3.4	1.13	1.15	80.90	19.10
Q004	0.61	0.067	7.3	7	2.22	2.4	84.70	15.30
Q005	0.68	0.065	19	17	1.72	2.53	95.50	4.50
Q006	0.88	0.065	2.1	2	2.52	2.57	90.60	9.40

finding that processing effectiveness depends on both the scale of geological partitioning and the degree of internal heterogeneity within each partition.

4.2. Implications for Adaptive SOM Implementation

Our findings have several practical implications for geoscientists applying UMLAs to geological interpretation.

1. Practitioners should reconsider conventional feature selection approaches that aggressively reduce dimensions to eliminate collinearity. As demonstrated in our experiments and supported by studies such as Xu et al. (2025b), collinear features can reinforce important geological patterns when used with appropriate clustering algorithms, provided the algorithm is configured with a sufficiently detailed topology and trained to convergence with appropriate parameters.
2. Our results support a layer-based approach where multiple geophysical transforms, even if mathematically derived from the same raw data, can substantially enhance boundary detection and lithological discrimination

10.1029/2025JH001107

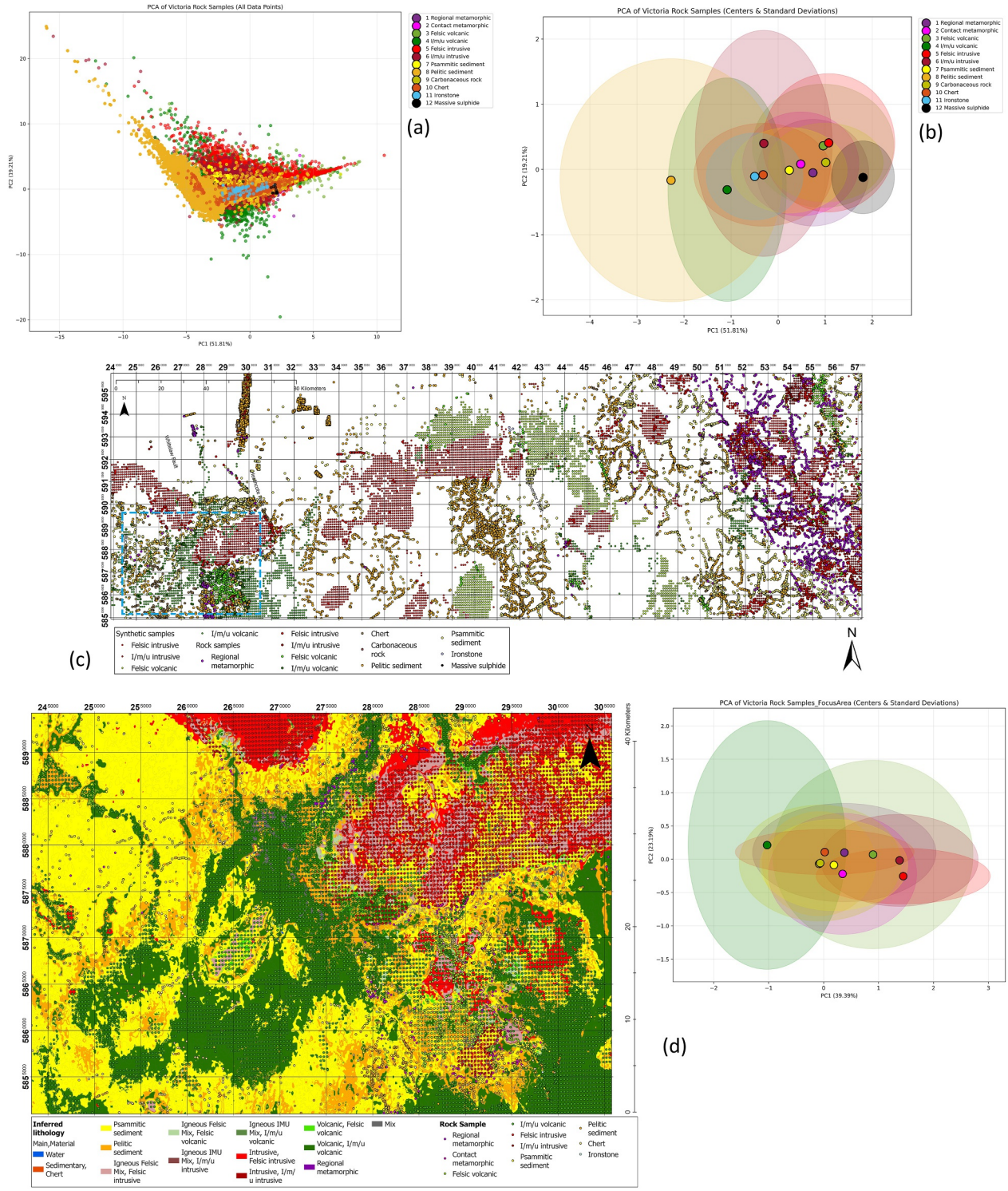

Figure 8. PCA of Victoria rock samples. (a) Scatter plot showing lithology types in PCA space (PC1 vs. PC2); (b) Elliptical representations of cluster means and standard deviations for each lithology; (c) Rock samples in Victoria. Modified from Xu et al. (2025b). The blue dashed box shows the location of the focused area in panel (c). The *x*-axis represents easting, unit in meters. The *y*-axis represents northing, unit in meters. Coordinate: GDA 94 (Zone 55). (c) Inferred lithology map of the focused area in Victoria, modified from Xu et al. (2025b), and its elliptical representations of cluster means and standard deviations for each lithology. The color presents the corresponding lithology units and sample units in the legends.

(Ghoneim et al., 2024; Wang et al., 2024). The N-Net's output layer appends class-probability dimensions to the original 8 geophysical features, yielding an $8 + n$ dimensional feature vector. Incorporating the learned class-probabilities is known to enhance class separability in deep neural networks (He & Su, 2023). Nonetheless, expanding the feature space reintroduces classic high-dimensional challenges. The curse of dimensionality (Banks & Fienberg, 2003) describes how adding dimensions exponentially inflates the feature-space volume. It also grows as the distance to the nearest neighbor approaches the distance to the farthest neighbor, meaning that distance-based metrics lose discriminative power (Beyer et al., 1999).

3. Our experiments demonstrate that additional improvements can be achieved through hybrid preprocessing strategies. The combination of dimension reduction techniques, for example, PCA, with post-processing approaches, for example, label smoothing (Gong et al., 2018; Müller et al., 2019), proved particularly effective, consistent with the findings of Nourani et al. (2013), who applied wavelet decomposition before SOM clustering. The development of adaptive SOM implementations should consider local collinearity patterns when determining optimal network architecture and training parameters. Areas with high input correlation may benefit from larger neighborhood functions to promote smoothing, while regions with independent features may require more focused learning to preserve sharp boundaries. This adaptive approach represents a departure from traditional uniform parameter selection and aligns with recent advances in spatially adaptive machine learning for geophysical applications.
4. Our study emphasizes that the optimal approach depends on the specific geological context and signal-to-noise characteristics of the available data. Research has consistently shown that derivative products improved edge detection despite introducing collinearity, while in low-noise settings lead to clear lithological contrasts (El-Omairi & El Garouani, 2023; Rafati et al., 2014; Zhenlong et al., 2023). This context dependence suggests that successful implementation requires careful consideration of local geological conditions rather than application of universal preprocessing protocols.

5. Conclusion

This study demonstrates that collinearity in geophysical data sets does not universally impair unsupervised machine learning classification performance, challenging conventional preprocessing assumptions. Through systematic evaluation across synthetic and real-world scenarios, we establish that Self-Organizing Maps and related distance-based clustering algorithms exhibit inherent robustness to feature correlation due to their mathematical formulation and topological optimization objectives. Our experimental framework reveals that collinear features improve classification when (a) cluster separation is minimal, requiring additional feature dimensions to resolve ambiguities, and (b) geological complexity demands enhanced boundary detection capabilities. The geospatial variability of collinearity effects necessitates adaptive rather than uniform preprocessing strategies, as optimal feature utilization depends on local geological complexity and signal characteristics.

This convergence reveals that collinearity is not merely a statistical artifact to be eliminated, but rather a signature of geological processes that can be used or mitigated depending on the context. For instance, alteration zones associated with mineralization often emerge in higher-order PCA components, which are frequently disregarded despite their relevance for mapping ore-related features (Behera & Panigrahi, 2022; Pradhan et al., 2022; Skilbrei & Kihle, 1999). Our results support adaptive layer-based approaches where multiple geophysical transforms, even if mathematically derived from the same raw data, can enhance boundary detection and lithological discrimination in complex geological settings while potentially introducing redundancy in simpler configurations.

While our synthetic experiments provide valuable controlled insights, several limitations warrant consideration. The synthetic models, while geologically realistic, represent idealized scenarios without the measurement noise, processing artifacts, and incomplete geological knowledge that characterize real-world data sets. The Victoria case study, though representative of southeastern Australian geology, captures a specific tectonic environment and lithological assemblage that may not encompass the full range of collinearity patterns present in other geological settings. Furthermore, quantitative performance metrics, while providing objective measures, may not fully capture the nuanced interpretation requirements that ultimately determine the success of clustering algorithms in supporting geological reasoning and hypothesis testing.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The magnetic, gravity, radiometric element, and geological map data supporting this study's findings are openly available at (Geological Survey of Victoria, 2008; Percival, 2014; Skladzien & Bibby, 2005; Thomas & Cudahy, 2012). Copyright belongs to © State of Victoria. The data is provided under the Digital Data Access Licence Agreement. Data processing and machine learning were performed using the following proprietary software:

1. MATLAB 2024a (The MathWorks Inc, 2024), accessed through the University of Melbourne Campus-Wide License.
2. ArcGIS Pro 3.3 (Esri Inc, 2023), accessed through the University of Melbourne Named User License.
3. Oasis Montaj (Seequent, 2023), accessed through the University of Melbourne Oasis Montaj Classroom Subscription.

A (https://doi.org/10.5281/zenodo.17493323) provides access to a GitHub repository associated with this manuscript (Xu et al., 2025a). The repository includes:

1. A simple synthetic model;
2. Six synthetic models constructed using Noddy;
3. A README.md file; and
4. A collection of Python scripts for sampling and unpacking the models.

All codes and data sets are released under the MIT License.

Acknowledgments
This work was supported by a PhD scholarship and a Faculty of Science Postgraduate Writing-Up Award from the University of Melbourne, awarded to LX. We acknowledge the exceptional efforts of the Geological Survey of Victoria in developing the open-source geological data set, which was fundamental to this research. The University of Melbourne made open-access publishing possible through the Council of Australian University Librarians agreement. Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australasian University Librarians.

References

Ahmad, A., & Dey, L. (2011). A K-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognition Letters*, *32*(7), 1062–1069. https://doi.org/10.1016/j.patrec.2011.02.017

Albuslimi, M., Alkalby, Y., & Al-Taweel, T. (2021). K-mean clustering analysis and logistic boosting regression for rock facies characterization and classification in Zubair Reservoir in Luhais Oil Field, Southern Iraq. *The Iraqi Geological Journal*, 65–75. https://doi.org/10.46717/igj.54.2B.6Ms-2021-08-26

Banks, D. L., & Fienberg, S. E. (2003). Data mining, statistics. In R. A. Meyers (Ed.), *Encyclopedia of physical science and technology* (3rd ed., pp. 247–261). Academic Press. https://doi.org/10.1016/B0-12-227410-5/00164-2

Bebis, G., & Georgiopoulos, M. (1994). Feed-forward neural networks. *IEEE Potentials*, *13*(4), 27–31. https://doi.org/10.1109/45.329294

Behera, S., & Panigrahi, M. K. (2022). Gold prospectivity mapping and exploration targeting in Hutti-Maski schist belt, India: Synergistic application of Weights-of-Evidence (Woe), Fuzzy Logic (Fl) and hybrid (Woe-Fl) models. *Journal of Geochemical Exploration*, *235*, 106963. https://doi.org/10.1016/j.gexplo.2022.106963

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "Nearest Neighbor" meaningful? Database theory—ICDT'99.

Boualla, O., Mehdi, K., & Zourarah, B. (2015). Collapse dolines susceptibility mapping in Doukkala Abda (Western Morocco) by using GIS matrix method (GMM). *Modeling Earth Systems and Environment*, *2*(1), 9. https://doi.org/10.1007/s40808-015-0064-8

Brazell, S., Bayeh, A., Ashby, M., & Burton, D. (2019). A machine-learning-based approach to assistive well-log correlation. *Petrophysics—The SPWLA Journal of Formation Evaluation and Reservoir Description*, *60*(4), 469–479. https://doi.org/10.30632/PJV60N4-2019a1

Cao, Y., Zhou, H., Yu, B., Wei, S., Chen, H., & Tian, Y. (2024). The estimation of petrophysical parameters based on ensemble smoother with correlation localization. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3403663

Carneiro, C. D. C., Fraser, S. J., Crósta, A. P., Silva, A. M., & Barros, C. E. D. M. (2012). Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon. *Geophysics*, *77*(4), K17–K24. https://doi.org/10.1190/geo2011-0302.1

Cooper, G. R. J., & Cowan, D. R. (2008). Edge enhancement of potential-field data using normalized statistics. *Geophysics*, *73*(3), H1–H4. https://doi.org/10.1190/1.2837309

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, *63*, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Curtis, S. M., & Ghosh, S. K. (2011). A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression. *Journal of Statistical Theory and Practice*, *5*(4), 715–735. https://doi.org/10.1080/15598608.2011.10483741

Dangeti, P. (2017). *Statistics for machine learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with python and R*. Packt Publishing Ltd.

Dascălu, C. G., & Cozma, C. D. (2009). The principal components analysis–method to reduce the collinearity in multiple linear regression model; application in medical studies. In *International conference on multivariate analysis and its application in science and engineering*.

De Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J., & Moguerza, J. M. (2022). General performance score for classification problems. *Applied Intelligence*, *52*(10), 12049–12063. https://doi.org/10.1007/s10489-021-03041-7

Dentith, M. C., & Mudge, S. (2014). *Geophysics for the mineral exploration geoscientist*. Cambridge University Press. https://doi.org/10.1017/CBO9781139024358

Doo, W.-B., Hsu, S.-K., & Yeh, Y.-C. (2007). A derivative-based interpretation approach to estimating source parameters of simple 2D magnetic sources from Euler deconvolution, the analytic-signal method and analytical expressions of the anomalies. *Geophysical Prospecting*, *55*(2), 255–264. https://doi.org/10.1111/j.1365-2478.2007.00603.x

Elahifar, B., & Hosseini, E. (2024). Automated real-time prediction of geological formation tops during drilling operations: An applied machine learning solution for the Norwegian Continental Shelf. *Journal of Petroleum Exploration and Production Technology*, *14*(6), 1661–1703. https://doi.org/10.1007/s13202-024-01789-5

El-Omairi, M. A., & El Garouani, A. (2023). A review on advancements in lithological mapping utilizing machine learning algorithms and remote sensing data. *Heliyon*, *9*(9), e20168. https://doi.org/10.1016/j.heliyon.2023.e20168

Esri Inc. (2023). Arcgis Pro. Version 3.3 [Software]. [Computer Program]. Retrieved from https://www.esri.com/en-us/arcgis/products/arcgis-pro/overviewhttps://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining, Portland, Oregon*.

Facchinelli, A., Sacchi, E., & Mallen, L. (2001). Multivariate statistical and gis-based approach to identify heavy metal sources in soils. *Environmental Pollution*, *114*(3), 313–324. https://doi.org/10.1016/S0269-7491(00)00243-8

Fairhead, J. D. (2015). *Advances in gravity and magnetic processing and interpretation*. European Association of Geoscientists and Engineers. https://doi.org/10.3997/9789462821774

Feyen, L., & Caers, J. (2006). Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. *Advances in Water Resources*, *29*(6), 912–929. https://doi.org/10.1016/j.advwatres.2005.08.002

Flexer, A. (1999). On the use of self-organizing maps for clustering and visualization. In *Principles of data mining and knowledge discovery*.

Ford, A., & Blenkinsop, T. G. (2008). Combining fractal analysis of mineral deposit clustering with weights of evidence to evaluate patterns of mineralization: Application to copper deposits of the Mount Isa Inlier, NW Queensland, Australia. *Ore Geology Reviews*, *33*(3), 435–450. https://doi.org/10.1016/j.oregeorev.2007.01.004

Fouedjio, F., & Klump, J. (2019). Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental Earth Sciences*, *78*(1), 38. https://doi.org/10.1007/s12665-018-8032-z

Geological Survey of Victoria. (2008). Statewide airborne magnetic and radiometric grids, 2005 [Dataset]. *Victoria—Gridded Airborne Geophysical Data, and Located and Gridded Gravity CD-Rom, updated 2008. Geological Survey of Victoria*. Retrieved from http://earthresources.efirst.com.au/product.asp?pID=22&cID=13

Ghezelbash, R., Maghsoudi, A., Shamekhi, M., Pradhan, B., & Daviran, M. (2023). Genetic algorithm to optimize the SVM and *K*-means algorithms for mapping of mineral prospectivity. *Neural Computing & Applications*, *35*(1), 719–733. https://doi.org/10.1007/s00521-022-07766-5

Ghoneim, S. M., Hamimi, Z., Abdelrahman, K., Khalifa, M. A., Shabban, M., & Abdelmaksoud, A. S. (2024). Machine learning and remote sensing-based lithological mapping of the Duwi Shear-Belt area, Central Eastern Desert, Egypt. *Scientific Reports*, *14*(1), 17010. https://doi.org/10.1038/s41598-024-66199-3

Golub, G. H., Hoffman, A., & Stewart, G. W. (1987). A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its Applications*, *88–89*, 317–327. https://doi.org/10.1016/0024-3795(87)90114-5

Gong, M. (2021). A novel performance measure for machine learning classification. *International Journal of Managing Information Technology*, *13*(1), 11–19. https://doi.org/10.5121/ijmit.2021.13101

Gong, W., Zhao, R., & Grünewald, S. (2018). Structured sparse *K*-means clustering via Laplacian smoothing. *Pattern Recognition Letters*, *112*, 63–69. https://doi.org/10.1016/j.patrec.2018.06.006

Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, *141*(1), 61–94. https://doi.org/10.1007/s11205-017-1832-9

Guo, J., Li, Y., Jessell, M. W., Giraud, J., Li, C., Wu, L., et al. (2021). 3D geological structure inversion from Noddy-generated magnetic data using deep learning methods. *Computers & Geosciences*, *149*, 104701. https://doi.org/10.1016/j.cageo.2021.104701

Hajnajafi, G., Jafarirad, A., Afzal, P., & Sheikh-Zakariaee, S.-J. (2021). Geological interpretation using multivariate K-means and robust factor analysis in Dezak area, SW Iran. *Environmental Earth Sciences*, *80*(1), 40. https://doi.org/10.1007/s12665-020-09305-8

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning* (Vol. 1, pp. 193–224). Springer. https://doi.org/10.1007/978-0-387-21606-5_7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), *The elements of statistical learning: Data mining, inference, and prediction* (pp. 485–585). Springer. https://doi.org/10.1007/978-0-387-84858-7_14

He, H., & Su, W. J. (2023). A law of data separation in deep learning. *Proceedings of the National Academy of Sciences of the USA*, *120*(36), e2221704120. https://doi.org/10.1073/pnas.2221704120

Holden, E.-J., Dentith, M., & Kovesi, P. (2008). Towards the automated analysis of regional aeromagnetic data to identify regions prospective for gold deposits. *Computers & Geosciences*, *34*(11), 1505–1513. https://doi.org/10.1016/j.cageo.2007.08.007

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Isles, D. J., & Rankin, L. R. (2013). Structural analysis from aeromagnetic data. In *Geological interpretation of aeromagnetic data* (pp. 63–93). Australian Society of Exploration Geophysicists. https://doi.org/10.1190/1.9781560803218.ch4

Jacobsen, B. H. (1987). A case for upward continuation as a standard separation filter for potential-field maps. *Geophysics*, *52*(8), 1138–1148. https://doi.org/10.1190/1.1442378

Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jessell, M. W., & Valenta, R. K. (1996). Structural geophysics: Integrated structural and geophysical modelling. In D. G. De Paor (Ed.), *Computer methods in the geosciences* (Vol. 15, pp. 303–324). Pergamon. https://doi.org/10.1016/S1874-561X(96)80027-7

Jolliffe, I. T. (2002). Principal components in regression analysis. In I. T. Jolliffe (Ed.), *Principal component analysis* (pp. 167–198). Springer. https://doi.org/10.1007/0-387-22440-8_8

Jooshaki, M., Nad, A., & Michaux, S. (2021). A systematic review on the application of machine learning in exploiting mineralogical data in mining and mineral industry. *Minerals*, *11*(8), 816. https://doi.org/10.3390/min11080816

Khalil, A., Abdel Hafeez, T. H., Saleh, H. S., & Mohamed, W. H. (2016). Inferring the subsurface basement depth and the structural trends as deduced from aeromagnetic data at West Beni Suef area, Western desert, Egypt. *NRIAG Journal of Astronomy and Geophysics*, *5*(2), 380–392. https://doi.org/10.1016/j.nrjag.2016.08.001

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International conference on learning representations*. https://doi.org/10.48550/arXiv.1412.6980

Kiviluoto, K. (1998). Predicting bankruptcies with the self-organizing map. *Neurocomputing*, *21*(1), 191–201. https://doi.org/10.1016/S0925-2312(98)00038-1

Kohonen, T. (1990). The self-organizing map. In *Proceedings of the IEEE* (Vol. 78(9), pp. 1464–1480). https://doi.org/10.1109/5.58325

Kohonen, T. (2001). The basic SOM. In T. Kohonen (Ed.), *Self-organizing maps* (pp. 105–176). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-56927-2_3

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Kohonen, T., Oja, E., Simula, O., Visa, A., & Kangas, J. (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE*, *84*(10), 1358–1384. https://doi.org/10.1109/5.537105

Li, Y., Wang, J., Zhao, H., Wang, C., & Shao, Q. (2023). Adaptive dbscan clustering and Gasa optimization for underdetermined mixing matrix estimation in fault diagnosis of reciprocating compressors. *Sensors*, *24*(1), 167. https://doi.org/10.3390/s24010167

Liang, B., Tang, C., Zhang, W., Xu, M., & Wu, T. (2023). N-Net: An unet architecture with dual encoder for medical image segmentation. *Signal, Image and Video Processing*, *17*(6), 3073–3081. https://doi.org/10.1007/s11760-023-02528-9

Lindsay, M. D., Perrouty, S., Jessell, M. W., & Aillères, L. (2013). Making the link between geological and geophysical uncertainty: Geodiversity in the Ashanti greenstone belt. *Geophysical Journal International*, *195*(2), 903–922. https://doi.org/10.1093/gji/ggt311

Liu, T., Yu, H., & Blair, R. H. (2022). Stability estimation for unsupervised clustering: A review. *WIREs Computational Statistics*, *14*(6), e1575. https://doi.org/10.1002/wics.1575

Lovie, P. (2005). Coefficient of variation. In *Encyclopedia of statistics in behavioral science*. https://doi.org/10.1002/0470013192.bsa107

Mason, C. H., & Perreault, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, *28*(3), 268–280. https://doi.org/10.1177/002224379102800302

Meigoony, M., Afzal, P., Gholinejad, M., Yasrebi, A., & Sadeghi, B. (2014). Delineation of geochemical anomalies using factor analysis and multifractal modeling based on stream sediments data in Sarajeh 1:100,000 sheet, central Iran. *Arabian Journal of Geosciences*, *7*(12), 5333–5343. https://doi.org/10.1007/s12517-013-1074-3

Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied Psychological Measurement*, *11*(4), 329–354. https://doi.org/10.1177/014662168701100401

Minty, B., Franklin, R., Milligan, P., Richardson, M., & Wilford, J. (2009). *The radiometric map of Australia (GA13928)*. Geoscience Australia. Retrieved from https://www.ga.gov.au/bigobj/GA13928.pdf

Miyamoto, S., Abe, R., Endo, Y., & Takeshita, J. I. (2015). Ward method of hierarchical clustering for non-euclidean similarity measures. In *2015 7th international conference of soft computing and pattern recognition (SoCPaR)*. https://doi.org/10.1109/SOCPAR.2015.7492784

Morlini, I. (2006). On multicollinearity and concurvity in some nonlinear multivariate models. *Statistical Methods and Applications*, *15*(1), 3–26. https://doi.org/10.1007/s10260-006-0005-9

Morrissey, M. B., & Ruxton, G. D. (2018). Multiple regression is not multiple regressions: The meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology*, *10*(3). https://doi.org/10.3998/ptpbio.16039257.0010.003

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*. https://doi.org/10.48550/arXiv.1906.02629

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, *31*(3), 274–295. https://doi.org/10.1007/s00357-014-9161-z

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2008). Handbook on constructing composite indicators and user guide (Vol. *2005*). https://doi.org/10.1787/533411815016

Nourani, V., Baghanam, A. H., Adamowski, J., & Gebremichael, M. (2013). Using self-organizing maps and wavelet transforms for space–time pre-processing of satellite precipitation and runoff data in neural network based rainfall–runoff modeling. *Journal of Hydrology*, *476*, 228–243. https://doi.org/10.1016/j.jhydrol.2012.10.054

Özdemir, S., Akbulut, Z., Karslı, F., & Acar, H. (2021). Automatic extraction of trees by using multiple return properties of the LiDAR point cloud. *International Journal of Electronic Governance*, *6*(1), 20–26. https://doi.org/10.26833/ijeg.668352

Palomino-Echeverria, S., Huergo, E., Ortega-Legarreta, A., Uson Raposo, E. M., Aguilar, F., Peña-Ramirez, C. D. L., et al. (2024). A robust clustering strategy for stratification unveils unique patient subgroups in acutely decompensated cirrhosis. *Journal of Translational Medicine*, *22*(1), 599. https://doi.org/10.1186/s12967-024-05386-2

Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: A review. *SIGKDD Explorations Newsletter*, *6*(1), 90–105. https://doi.org/10.1145/1007730.1007731

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(null), 2825–2830. https://doi.org/10.5555/1953048.2078195

Perafan-Lopez, J. C., Ferrer-Gregory, V. L., Nieto-Londoño, C., & Sierra-Pérez, J. (2022). Performance analysis and architecture of a clustering hybrid algorithm called Fa+Ga-Dbscan using artificial datasets. *Entropy*, *24*(7). https://doi.org/10.3390/e24070875

Percival, P. J. (2014). Digital files for the index of airborne geophysical surveys version 14) [Dataset]. *Geoscience Australia*. Retrieved from https://pid.geoscience.gov.au/dataset/ga/79135, https://researchdata.edu.au/digital-files-index-edition-2014

Portales, L., Cazelles, E., & Pauwels, E. (2025). On the sequential convergence of Lloyd's algorithms. *Mathematics of Operations Research*. https://doi.org/10.1287/moor.2024.0550

Pradhan, B., Jena, R., Talukdar, D., Mohanty, M., Sahu, B. K., Raul, A. K., & Abdul Maulud, K. N. (2022). A new method to evaluate gold mineralisation-potential mapping using deep learning and an explainable artificial intelligence (Xai) model. *Remote Sensing*, *14*(18), 4486. https://doi.org/10.3390/rs14184486

Rafati, M., Arabfard, M., & Rafati-Rahimzadeh, M. (2014). Comparison of different edge detections and noise reduction on ultrasound images of carotid and brachial arteries using a speckle reducing anisotropic diffusion filter. *Iranian Red Crescent Medical Journal*, *16*(9), e14658. https://doi.org/10.5812/ircmj.14658

Ramos Emmendorfer, L., & de Paula Canuto, A. M. (2021). A generalized average linkage criterion for hierarchical agglomerative clustering. *Applied Soft Computing*, *100*, 106990. https://doi.org/10.1016/j.asoc.2020.106990

Reading, A. M., Cracknell, M. J., Bombardieri, D. J., & Chalke, T. (2015). Combining machine learning and geophysical inversion for applied geophysics. *ASEG Extended Abstracts*, *2015*(1), 1–5. https://doi.org/10.1071/ASEG2015AB070

Reid, A. B. (2007). Euler deconvolution. In D. Gubbins & E. Herrero-Bervera (Eds.), *Encyclopedia of geomagnetism and paleomagnetism* (pp. 263–265). Springer. https://doi.org/10.1007/978-1-4020-4423-6_98

Reid, A. B., Allsop, J. M., Granser, H., Millett, A. J., & Somerton, I. W. (1990). Magnetic interpretation in three dimensions using Euler deconvolution. *Geophysics*, *55*(1), 80–91. https://doi.org/10.1190/1.1442774

Reynolds, D. (2009). Gaussian mixture models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of biometrics* (pp. 659–663). Springer. https://doi.org/10.1007/978-0-387-73003-5_196

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sammut, C., & Webb, G. I. (2010). Expectation-maximization algorithm. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning*. Springer. https://doi.org/10.1007/978-0-387-30164-8_291.387.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., et al. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Seequent. (2023). Oasis Montaj. In Version 2024.1 [Software]. [Computer Program]. Retrieved from https://www.seequent.com/products-solutions/geosoft-oasis-montaj/https://www.seequent.com/products-solutions/geosoft-oasis-montaj/

Shahrestani, S., & Sanislav, I. (2025). Delineation of geochemical anomalies through empirical cumulative distribution function for mineral exploration. *Journal of Geochemical Exploration*, *270*, 107662. https://doi.org/10.1016/j.gexplo.2024.107662

Skilbrei, J. R., & Kihle, O. (1999). Display of residual profiles versus gridded image data in aeromagnetic study of sedimentary basins: A case history. *Geophysics*, *64*(6), 1740–1747. https://doi.org/10.1190/1.1444679

Skladzien, P. B., & Bibby, D. (2005). Geophysics radiometric statewide images In: Radiometric grid of Australia (Radmap)—Ternary image (K, Th, U) [Dataset]. *Geological Survey of Victoria*. Retrieved from http://earthresources.efirst.com.au/product.asp?pID=22&cID=13

Smith, L., Horrocks, T., Holden, E.-J., Wedge, D., & Akhtar, N. (2022). Magnetic grid resolution enhancement using machine learning: A case study from the eastern goldfields superterrane. *Ore Geology Reviews*, *150*, 105119. https://doi.org/10.1016/j.oregeorev.2022.105119

Smith, S. P., & Dubes, R. (1980). Stability of a hierarchical clustering. *Pattern Recognition*, *12*(3), 177–187. https://doi.org/10.1016/0031-3203(80)90042-4

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. https://doi.org/10.5555/2627435.2670313

Stavrev, P., & Reid, A. (2007). Degrees of homogeneity of potential fields and structural indices of Euler deconvolution. *Geophysics*, *72*(1), L1–L12. https://doi.org/10.1190/1.2400010

The MathWorks Inc. (2024). Matlab. Version 2024a [Software]. [Computer Program]. Retrieved from https://www.mathworks.com/products/matlab.htmlhttps://www.mathworks.com/products/matlab.html

Thomas, M., & Cudahy, T. (2012). National Aster geoscience map. G. Australia [Dataset]. Retrieved from https://pid.geoscience.gov.au/dataset/ga/74427

Thompson, D. T. (1982). Euldph: A new technique for making computer-assisted depth estimates from magnetic data. *Geophysics*, *47*(1), 31–37. https://doi.org/10.1190/1.1441278

Todaro, V., D'Oria, M., Tanda, M. G., & Gómez-Hernández, J. J. (2021). Ensemble smoother with multiple data assimilation to simultaneously estimate the source location and the release history of a contaminant spill in an aquifer. *Journal of Hydrology*, *598*, 126215. https://doi.org/10.1016/j.jhydrol.2021.126215

Tokuda, E. K., Comin, C. H., & Costa, L. D. F. (2022). Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and its Applications*, *585*, 126433. https://doi.org/10.1016/j.physa.2021.126433

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249–267. https://doi.org/10.1016/j.wocn.2018.09.004

Uieda, L., & Barbosa, V. (2012). Robust 3D gravity gradient inversion by planting anomalous densities. *Geophysics*, *77*(4), G55–G66. https://doi.org/10.1190/GEO2011-0388.1

Vettigl, G. (2025). *Minisom: Minimalistic implementation of self-organizing maps* [Code]. [Online Database]. Distributor. Retrieved from https://github.com/JustGlowing/minisom

Wallet, B. C., & Hardisty, R. (2019). Unsupervised seismic facies using Gaussian mixture models. *Interpretation*, *7*(3), SE93–SE111. https://doi.org/10.1190/int-2018-0119.1

Wang, W., Xue, C., Zhao, J., Yuan, C., & Tang, J. (2024). Machine learning-based field geological mapping: A new exploration of geological survey data acquisition strategy. *Ore Geology Reviews*, *166*, 105959. https://doi.org/10.1016/j.oregeorev.2024.105959

Wilks, D. S. (2019). Chapter 13—Principal component (EOF) analysis. In D. S. Wilks (Ed.), *Statistical methods in the atmospheric sciences* (4th ed., pp. 617–668). Elsevier. https://doi.org/10.1016/B978-0-12-815823-4.00013-4

Willman, C. E., Cayley, R. A., VandenBerg, A. H. M., Haydon, S. J., Osborne, C. R., Seymon, A. R., & Thom, J. L. (2005). *Dargo 1:100 000 Map Area Geological Report*. Geological Survey of Victoria Report 126) GeoScience Victoria.

Wynne, P., & Bacchin, M. (2009). *Index of gravity surveys* (2nd edn) (Record No. 2009/07) [Online Database]. Distributor. Retrieved from http://www.ga.gov.au/gadds

Xu, L., & Green, E. C. R. (2023). Inferring geological structural features from geophysical and geological mapping data using machine learning algorithms. *Geophysical Prospecting*, *71*(9), 1728–1742. https://doi.org/10.1111/1365-2478.13371

Xu, L., Green, E. C. R., & Feltrin, L. (2025a). Rethinking collinearity in self-organising maps: Evidence from Geophysical Data Classification—Code (1.0). In (Version 1.0.0). [Computer Program]. *Zenodo*. https://doi.org/10.5281/zenodo.17493323

Xu, L., Green, E. C. R., & Kelly, C. (2024). Inferring fault structures and overburden depth in 3D from geophysical data using machine learning algorithms—A case study on the Fenelon gold deposit, Quebec, Canada. *Geophysical Prospecting*, *72*(9), 3474–3494. https://doi.org/10.1111/1365-2478.13589

Xu, L., Green, E. C. R., McLean, M. A., & Feltrin, L. (2025b). Applying SOM cluster analysis with iterative refinement to infer lithology units in eastern Victoria. *Earth and Space Science*, *12*(8), e2024EA003999. https://doi.org/10.1029/2024EA003999

Xu, L., Green, E. C. R., McLean, M. A., & Feltrin, L. (2025c). Applying SOM cluster analysis with iterative refinement to infer lithology units in eastern Victoria, Australia—Code (1.0). In (version 1.0.0). [Computer Program]. *Zenodo*. https://doi.org/10.5281/zenodo.15502075

Yin, H. (2008). The self-organizing maps: Background, theories, extensions and applications. In J. Fulcher & L. C. Jain (Eds.), *Computational intelligence: A compendium* (pp. 715–762). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78293-3_17

Zhenlong, H., Jikang, W., Jinrong, S., Xinwei, L., & Wentian, Z. (2023). Intelligent lithology identification methods for rock images based on object detection. *Natural Resources Research*, *32*(6), 2965–2980. https://doi.org/10.1007/s11053-023-10271-8