Machine Learning Across Data Domains: Methods and Applications in Geological Analysis

av

Suraj Neelakantan

Akademisk avhandling

Avhandling för teknologi doktorsexamen i datavetenskap,
som kommer att försvaras offentligt
torsdag den 12 november 2026 kl. 13:15,
Hörsal L2, Örebro universitet

Opponent: Associate Professor Foteini Liwicki
Luleå University of Technology
Luleå, Sverige

Örebro universitet
Institutionen för naturvetenskap och teknik
701 82 ÖREBRO

Abstract

Machine learning and deep learning have become central tools for automating analysis tasks across scientific domains, yet their effectiveness heavily depends on assumptions that rarely hold outside curated benchmarks: that labeled training data is plentiful and unambiguous, that input measurements come from a single homogeneous source, and that model outputs need not be adjusted by the user. When these assumptions break down, as they routinely do in real-world scenarios, standard methods produce models that are either not accurate or difficult to trust. Geology is one such field where many of these assumptions fail. Annotations are expensive and subjective, with trained experts frequently disagreeing on the same sample. Instruments such as X-ray computed tomography and X-ray fluorescence scanners capture complementary but structurally different measurements of the same material. In other cases, domain experts might need to inspect and sometimes override outputs generated by machine learning models before acting on them. These are not uniquely geological problems; they reflect broader methodological gaps in how machine learning handles label scarcity, data heterogeneity, and user interaction. The research presented here addresses three such gaps, using geological datasets as the primary experimental setting.

The first research direction is on the distributional mismatch between synthetic and real data for semantic segmentation. A pipeline for generating synthetic training data from known physical parameters is combined with adversarial domain adaptation, implemented through gradient reversal within a U-Net architecture, to transfer segmentation models from synthetic X-ray computed tomography images to real sensor scans. Four learning configurations sharing an identical backbone isolate the effect of adaptation from architectural differences. The semi-supervised variant achieves performance comparable to a fully supervised model while drawing most of its supervision from automatically labeled synthetic data. Notably, domain-adapted models approach the level of agreement observed among independent human annotators, suggesting that annotation subjectivity, rather than model capacity, bounds achievable performance.

The second research direction addresses the integration of measurements from multiple geological data types for lithology classification. Feature-level fusion of elemental concentrations from X-ray fluorescence with attenuation statistics from X-ray computed tomography is evaluated across several model families, including tree-based ensembles, Bayesian neural networks, and attentionbased tabular architectures. Integration improves classification accuracy, with the most pronounced gains at boundaries between classes that share similar signatures in one measurement domain but differ in the other. A separate but related contribution introduces a framework in which large language models encode domain constraints into data augmentation policies expressed in natural language. The resulting synthetic samples respect the compositional structure that geometry-based methods such as SMOTE disregard, achieving stronger minority-class performance with fewer generated examples.

The final research direction investigates whether image super-resolution can achieve both interpretability and controllability without compromising reconstruction quality. A dual-branch network separately learns fine textures and larger structures, then combines them using channel and spatial attention. By introducing an adjustable parameter at inference time that controls the relative contribution of each branch, the model enables dynamic navigation of the perception–distortion tradeoff without requiring retraining. Pairwise preference studies further indicate that human evaluators consistently prefer outputs generated with tuned branch weights over those produced using fixed configurations.

Together, the three directions demonstrate that adapting the machine learning setup to the specific constraints of the application area, such as multiple data domains, the scarcity of labels, and the rigidity of model outputs, produces models that are more accurate, more interpretable, and better aligned with the needs of domain practitioners. The experimental setting is geological, but the methods and findings address limitations common to applied machine learning in any scientific domain where these constraints are present.