Scientific Contributions Oil & Gas, Vol. 49. No. 1, March: 375 - 393

SCIENTIFIC CONTRIBUTIONS OIL AND GAS

Testing Center for Oil and Gas
LEMIGAS

Journal Homepage:http://journal.lemigas.esdm.go.id
ISSN: 2089-3361, e-ISSN: 2541-0520

A Conceptual Framework for Cross-Basin Reservoir Characterization: Integrating Unsupervised-Supervised Learning to Address Geological Heterogeneity

Rian Cahya Rohmana [1,2] , Tutuka Ariadji [1], Amega Yasutra[1], Dedy Irawan[1]

[1]Department of Petroleum Engineering, Institut Teknologi Bandung
Ganesha Street No.10 Bandung, Indonesia.

[2]Petroleum Engineering, Tanri Abeng University
Swadharma Raya Street No.58 Jakarta, Indonesia.

Corresponding Author : Tutuka Ariadji (tutukaariadji@itb.ac.id)

Manuscript received: January 5th, 2025; Revised: January 26th, 2026
Approved: January 27th, 2026; Available online: March 16th, 2026; Published: March 16th, 2026.

ABSTRACT Machine Learning (ML) offers an innovative approach to reservoir characterization, yet its widespread application is hindered by poor cross-basin generalization. Specifically, most models validated on intra-basin data fail to account for the geological heterogeneity encountered in new settings. This study addresses this limitation by quantifying the geological domain shift that hinders direct model transferability and by proposing a conceptual framework to overcome it. This approach is designed to systematically address geological variability by identifying local electrofacies structures before attempting predictive modelling. To validate the necessity of this framework, we conducted a preliminary empirical study comparing the Talang Akar Formation in the North West Java Basin and the Menggala Formation in the Central Sumatera Basin (Basins A and B). This study empirically demonstrates that, despite shared fluvial-deltaic depositional environments, the intrinsic statistical structures of the two basins diverge significantly. A quantitative analysis revealed a severe domain shift, evidenced by a large Euclidean distance between cluster centroids and a low adjusted Rand index (ARI) of 0.113 when applying direct analog mapping. These findings empirically demonstrate that direct model transfer is ineffective because of second-order geological controls. Consequently, this study establishes the critical need for the proposed UL-SL strategy to adaptively handle domain shifts, providing a geologically grounded roadmap for accurate characterization in frontier and data-scarce basins.

Keywords: machine learning, reservoir characterization, cross-basin validation, geological heterogeneity.

How to cite this article:

Rian Cahya Rohmana, Tutuka Ariadji, Amega Yasutra and Dedy Irawan 2026, A Conceptual Framework For Cross-Basin Reservoir Characterization: Integrating Unsupervised-Supervised Learning To Address Geological Heterogeneity, Scientific Contributions Oil and Gas, 49 (1) pp. 375-393. DOI org/10.29017/scog.v49i1.2001

INTRODUCTION

The precise characterization of subsurface reservoirs is crucial for the exploration and development of hydrocarbon fields. Petrophysical properties, including rock facies, porosity (ϕ), permeability (k), and water saturation (Sw), are used for reserve estimation, production planning, and reservoir management. These properties are traditionally determined by analyzing rock cores and interpreting well logs. Core analysis yields accurate direct measurements but is expensive and provides data from limited intervals in sparse wells. Conversely, well log interpretation provides continuous vertical data but can be time-consuming and subjective. Furthermore, conventional log interpretation often fails to capture the complex nonlinear relationships between log responses and rock properties, particularly in heterogeneous reservoirs (Dwihusna, 2020; Pan, 2022; Serra, 1984).

To address these limitations, the industry has adopted data-driven methods, particularly machine learning (ML) and deep learning (DL). ML and DL models can identify complex nonlinear patterns in large multi-attribute datasets more rapidly and objectively than conventional methods. Accordingly, numerous studies have used algorithms, such as support vector machines (SVM), random forests (RF), and various neural network (NN) architectures (e.g., convolutional neural networks [CNN] and long short-term memory [LSTM]), to predict electrofacies (Dwihusna, 2020; Gu et al., 2019; He et al., 2020; Imamverdiyev & Sukhostat, 2019; Li et al., 2020; Pratama, 2018; Singh et al., 2020; Verma et al., 2021; Candra et al., 2024), porosity, permeability, and elastic properties (Iraji et al., 2023; Wood, 2020; Ulil et al., 2025), and water saturation.

A primary challenge limiting the widespread application of ML and DL models in petrophysics is their poor generalization ability for new datasets. Most published studies use intra-basin validation, in which models are tested on holdout data from the same field or geological basin as the training set (Pratama, 2018; Singh et al., 2020; Wood, 2020). Although this approach confirms model performance within a single geological domain, it fails to address the utility of models in novel areas. The challenge of model generalization is not new; in fact, early studies applying neural networks in the late 1990s had already explicitly identified the local dependence of these models, warning that they are less likely to be successful when applied elsewhere (Gonçalves et al., 1997).

The ability of these models to generalize across different geological basins, a process known as cross-basin validation, remains largely unverified in the literature. This challenge arises because the relationship between well-log responses and petrophysical properties is not universal; it is controlled by the geological context of each basin, including its depositional history, tectonics, and diagenesis. This geological heterogeneity causes a shift in the data distribution between basins (Li et al., 2021; Yang et al., 2023; Zainuri et al., 2023). Consequently, a model trained in one basin is likely to show poor performance when applied to another basin (Dramsch, 2020). This limitation prevents the use of ML models for scalable regional exploration or the evaluation of frontier basins where data are sparse (Brackenridge et al., 2022), a challenge that is conceptually illustrated by an iceberg analogy (Figure 1).

This study addresses the problem of poor cross-basin generalization for machine learning (ML) models in petrophysics. To do so, we:

- The current use of intra-basin validation in the ML petrophysics literature was reviewed to establish the limitations of this approach.
- Analyze the effects of geological factors, such as depositional environment and diagenesis, on

data distribution shifts that limit model generalization.

- A conceptual framework is proposed to improve the generalization. This framework integrates two components: (a) geologically constrained validation using formations with analogous depositional environments and (b) a combined unssupervised–supervised learning (UL-SL) strategy that uses electrofacies to guide predictions across basins.

METHODOLOGY

LITERATURE REVIEW

Machine learning in reservoir characterization

The application of machine learning (ML) using oil and gas data has advanced algorithmically; however, the methods for model validation have not kept up. While computational methods in geosciences date back over 70 years (Figure 2) (Dramsch, 2020), the 1990s marked a shift from knowledge-based systems to data-driven methods, such as support vector machines (SVMs) and random forests (RFs) (Cortes and Vapnik, 1995; Ho, 1995 in Dramsch, 2020). Recent advances in parallel computing and open-source libraries have accelerated the adoption of these algorithms. Numerous studies have utilized RF, SVM, and deep learning architectures to predict permeability (Khan et al., 2019; Talebkeikhah et al., 2021), porosity (Wood, 2020), and classify electrofacies (Singh et al., 2020; Verma et al., 2021). These modern approaches build upon foundational work that used artificial neural networks as model-free function estimators to overcome the limitations of simple regression (Mohaghegh et al., 1995, 1997).

However, despite these algorithmic advances, a review of the literature reveals a consistent reliance on intra-basin validation methods. Model performance is typically assessed using techniques such as holdout sets or k-fold cross-validation, in which all data for training and testing are sourced from the same geological basin (Khan et al., 2019; Nugroho et al., 2024; Pratama, 2018; Wood, 2020; Zhang et al., 2021).

This practice is problematic because it relies on the assumption that the training and test data are drawn from the same distribution. Although valid for local applications, this assumption fails for different geological basins. Consequently, the majority of reviewed studies rely exclusively on this technique, underscoring a significant gap in

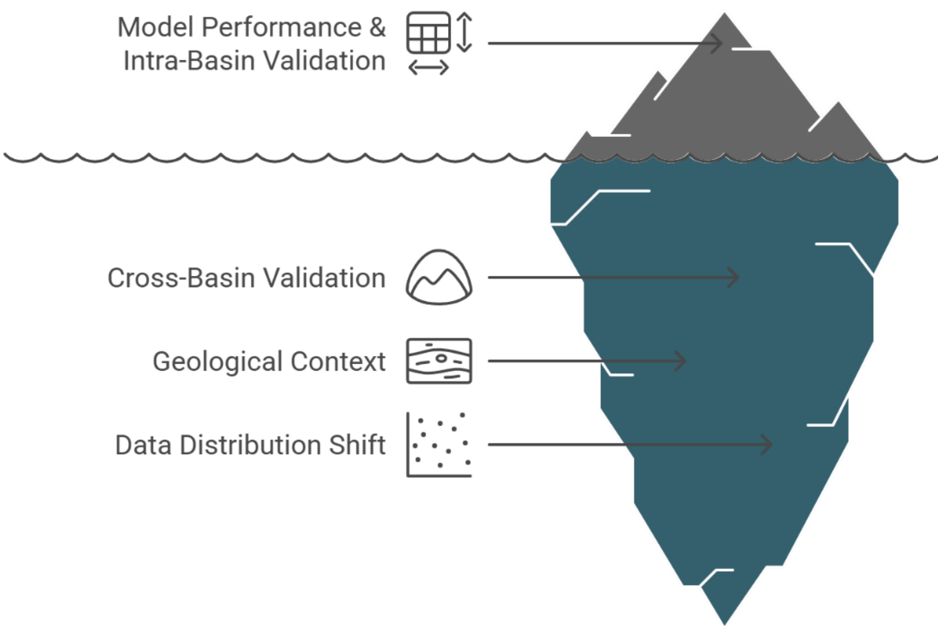

Figure 1 Iceberg analogy for machine learning model validation. Whereas conventional intra-basin performance metrics represent the visible tip, the more critical submerged portion represents the fundamental barriers to generalizing geological context, cross-basin data shifts, and the need for more validation protocols. The figure suggests that true model robustness depends on addressing these hidden geological factors, not only on optimizing the visible metrics.

verifying the true generalizability of these models for regional and exploratory applications in the future.

Early researchers proposed a classify-then-predict paradigm to address geological heterogeneity by partitioning data into meaningful groups, such as lithofacies, before prediction (Lee & Datta-Gupta, 1999; Wong et al., 1995). Recently, transfer learning (TL) and domain adaptation (DA) have been applied to account for data distribution shifts between basins (Li et al., 2021; Yang et al., 2023). TL strategies, such as pre -training and fine-tuning, have been shown to accelerate modeling and reduce data requirements (Dong et al., 2021; Kompantsev, 2024; Misra et al., 2024). Similarly, DA methods attempt to align the statistical distributions of the source and target domains (Yao et al., 2020).

However, these approaches have significant limitations in quantitative petrophysical predictions. First, they did not systematically address the challenge of predicting static petrophysical properties from well logs in geologically distinct basins. Second, these methods often function as black boxes that align distributions through complex mathematical transformations. This lack of interpretability is a disadvantage in geological data, where understanding the physical basis of a prediction is essential (Cuddy, 2000). A research gap remains for a framework that can systematically address cross-basin prediction in a manner that is both geologically transparent and interpretable. To bridge this gap, the proposed UL–SL framework addresses this need by using unsupervised learning to identify geologically significant electrofacies and find their analogs across basins, thereby shifting the paradigm from purely statistical correction to geologically informed adaptation.

Geological heterogeneity

The poor generalization of machine learning models across different basins is a direct consequence of geological heterogeneity rather than algorithmic deficiency. Each sedimentary basin possesses a unique geological history that governs the relationship between rock properties and well-log responses. This basin-scale heterogeneity results in a data distribution shift that violates the assumption that the training and testing data originate from the same distribution (Li et al., 2021; Yang et al., 2023). Consequently, a model trained on data from one basin is forced to extrapolate beyond its learned domain when applied to another basin, leading to inaccurate predictions (Dramsch, 2020).

The depositional environment serves as the primary control on the initial reservoir properties, including the facies and pore network architecture (Leila & Moscariello, 2018; Lorenz et al., 1989; Merletti et al., 2017; Omoboriowo et al., 2012). Distinct facies, such as fluvial channels or marine mud, produce predictable signatures in well-log data (Abdel-Fattah, 2015). This principle underpins the use of geological analogies, in which formations sharing a common depositional origin provide a valid baseline for comparison (Ismail et al., 2025). However, this initial rock fabric is subsequently altered by post-depositional processes that are unique to the tectonic and burial histories of each basin (Sarhan, 2022; Smeraglia et al., 2014).

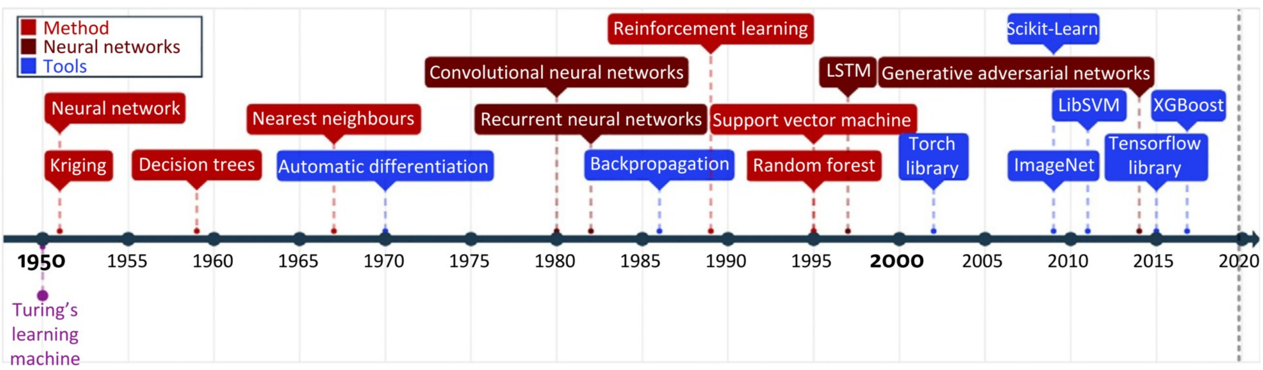

Figure 2 Machine learning timeline (Dramsch, 2020).

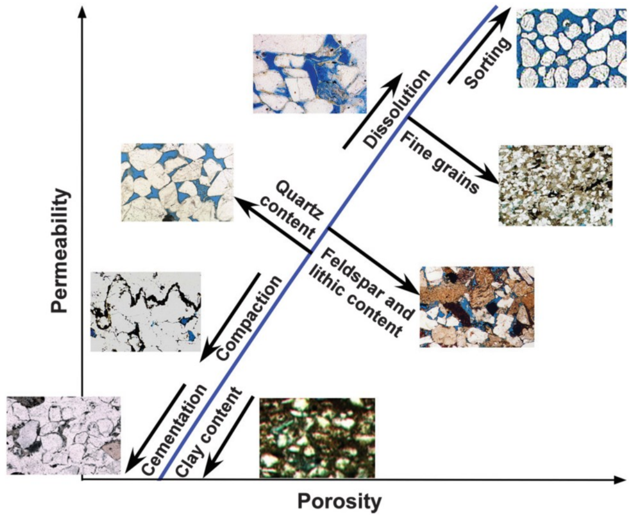

Figure 3 Depositional and diagenetic controls on the porosity-permeability relationship in clastic reservoirs. (Merletti et al., 2017).

Diagenetic processes involving physical and chemical changes fundamentally modify rocks and their pore systems, thereby affecting reservoir quality. Compaction and the precipitation of minerals, such as quartz or calcite, reduce primary porosity and can drastically lower permeability (Jafari et al., 2020; Magoba et al., 2024). Conversely, corrosive fluids can dissolve grains to create secondary porosity that enhances reservoir quality, independent of the original fabric (Lima et al., 2022; Rashid et al., 2022).

As illustrated in Figure 3, these competing processes create a diagenetic overprint that decouples the final petrophysical properties from the initial depositional trend. This explains why rocks from different basins are not directly comparable, even when they share a depositional origin. A gamma ray value indicating high permeability in a minimally cemented basin may correspond to tight rock in a basin that has undergone intense diagenesis. Therefore, a successful cross-basin framework must account for data distribution shifts caused by the unique geological histories of the basins.

Proposed framework for cross-Basin generalization

To address the domain shift quantified in this study, we propose a conceptual framework designed to bridge the gap between geological analogies and statistical compatibility. This framework shifts the paradigm from simple direct prediction to a geologically adaptive workflow that accounts for the distinct rock properties inherent to each basin. The framework is based on two primary pillars: geologically constrained validation and an integrated unsupervised–supervised learning strategy. The first component dictates that the training data must be strictly selected from formations sharing analogous depositional environments. This step is essential to control for primary geological variability and ensure that the model starts from a shared sedimentary template. By doing this, we effectively isolate the effects of secondary overprints, such as diagenesis and tectonics, which allows the subsequent algorithmic steps to focus on adjusting for these specific local factors.

The core algorithmic innovation is an integrated unsupervised–supervised learning strategy. This workflow is designed to account for statistical shifts that occur even between analog formations by allowing the target data to inform the model structure before any predictions are made. Ideally, the workflow should proceed in a specific sequence of operations.

First, an unsupervised algorithm was applied independently to the well log data from both the source and target basins. This step identified the distinct electrofacies or local data structures inherent to each basin without external bias. The objective was to capture the unique statistical signature of the rock, such as the high-density clusters observed in our study, without forcing them into a pre-existing category.

In this study, the unsupervised stage applies centroid-based clustering (K-means) to standard well logs to derive electrofacies as a rock-type proxy. Each depth sample receives a discrete electrofacies label. The clustering step also provides centroids and within-cluster dispersion, which summarize the dominant log-response patterns in each basin.

In the supervised stage, a predictive model is trained on source basin data to estimate the target variable. Based on label availability, the task is formulated as regression, such as petrophysical property prediction, or classification, such as electrofacies or rock-type labeling. The key integration step is electrofacies-conditioned supervision. This is implemented by using cluster labels as categorical inputs, fitting separate supervised models for each electrofacies, or applying optional feature alignment before training to mitigate cross-basin shift.

Consider an example in which clustering separates intervals into clean sand, shaly sand, and tight zones. Under this structure, a single global regressor for porosity or permeability can be suboptimal because each electrofacies may follow a different rock-physics relationship. An electrofacies-specific strategy instead trains separate regressors within source basin clusters. During target basin prediction, target clusters are first identified and then mapped to the most similar source clusters using centroid similarity. The mapped electrofacies label is then used to select the corresponding electrofacies-conditioned predictor, preventing the application of a clean-sand relationship to shale-dominated intervals.

Second, the framework mandates a mapping process in which electrofacies from the target basin are aligned with those from the source basin based on their statistical properties. Unlike rigid classification, this mapping acts as a translation layer that correlates the unique clusters of the new basin with the known facies of the training basin.

Finally, the resulting cluster labels from the source basin were included as categorical features in the training dataset for the supervised model. When applying the model to the target basin, the mapped electrofacies labels were used as input features alongside the standard logs. This approach effectively provides a supervised model with a geological context or prior belief regarding the data structure of the target basin.

By explicitly feeding the model information about the local cluster distribution, the algorithm can adapt its internal weights to accommodate local shifts in mineralogy or compaction. This theoretical approach ensures that the model is not blindly extrapolated but instead makes informed predictions based on the actual statistical reality of the new environment.

This study proposes a hybrid unsupervised–supervised pipeline. First, electrofacies are derived by independent clustering in the source and target basins. Next, target electrofacies are mapped to source electrofacies to form a translation layer between basins. The supervised model is trained in the source basin using standard logs and source electrofacies labels as categorical features. For target basin prediction, the model uses mapped electrofacies labels as additional inputs.. Figure 4 summarizes the proposed workflow.

RESULT AND DISCUSSION

To validate the necessity of the proposed unsupervised-supervised learning (UL-SL) framework, we conducted a controlled empirical study using real-world datasets from two distinct Indonesian hydrocarbon basins. The objective of this study was not merely to test model performance but to quantify the statistical divergence between formations that are conceptually considered geological analogs. This section presents empirical evidence of the domain shift, which acts as a fundamental barrier to cross-basin generalization.

Experimental setup and geological context

This study utilized datasets from two prolific basins (Figure 5), designated as the source (Basin A) and target (Basin B) domains. The source

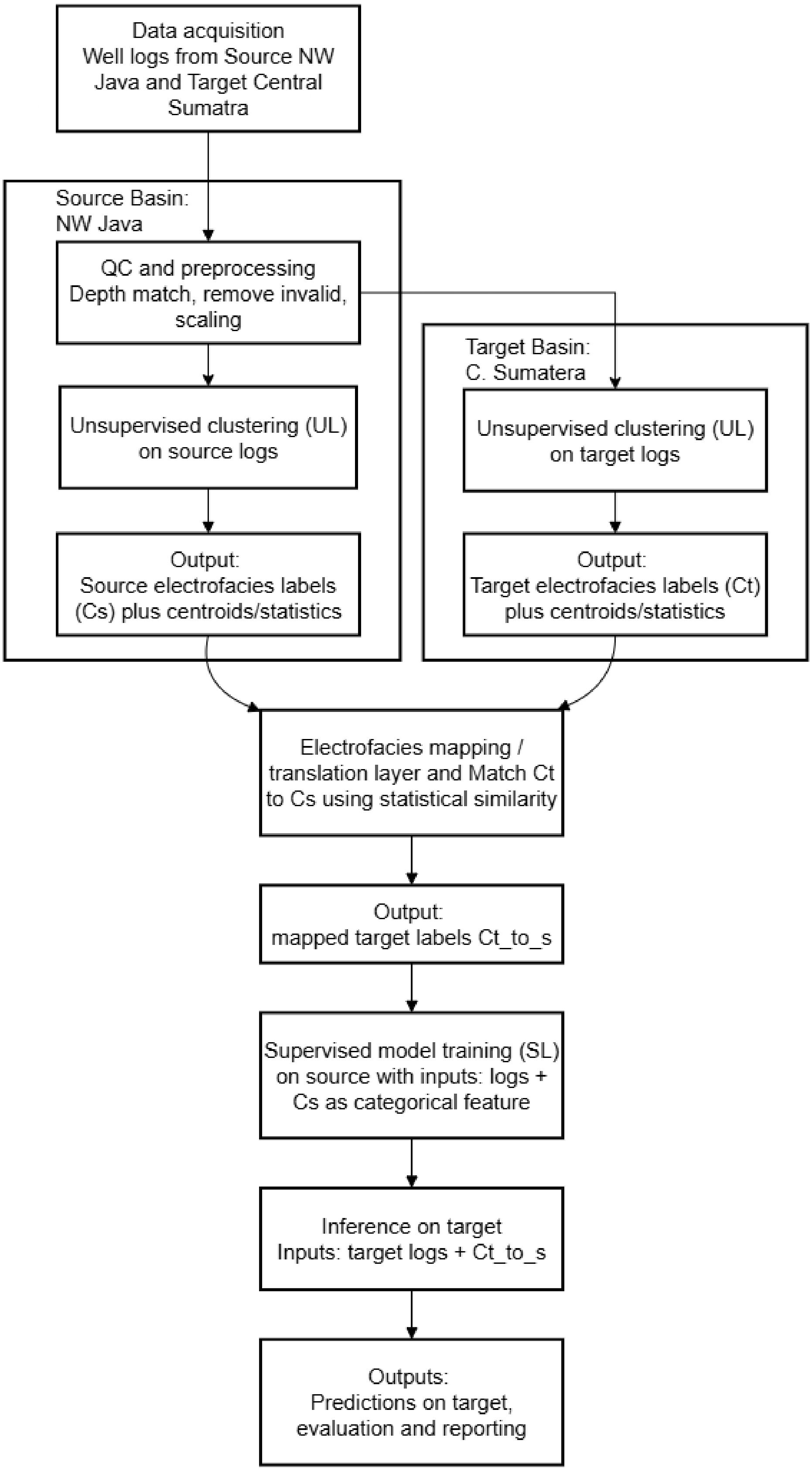

Figure 4. Proposed workflow

dataset comprised well-log data from 25 wells penetrating the Talang Akar Formation in the North West Java Basin. Geologically, the Talang Akar Formation was deposited in transitional settings ranging from fluvial to deltaic and shallow-marine environments. It is characterized by complex lithological heterogeneity, consisting of intercalated sandstone reservoirs, claystone, siltstone, and localized thin coal layers. The target dataset (seven wells) was derived from the Menggala Formation in the Central Sumatra Basin. Similar to the Talang Akar Formation, the Menggala Formation serves as a primary reservoir and is deposited in comparable fluvial-deltaic environments.

These two formations were selected because they represent analogous depositional environments in fluvial deltaic systems, allowing depositional settings to function as a controlled baseline in the experimental design. By maintaining the primary depositional origin broadly constant, the study focuses on inter-basins that may drive domain shift in well-log responses.

These differences include tectonic setting and activity, sediment provenance with associated mineralogical trends, and basin-specific diagenetic overprints. One important contrast is that the NW Java Basin stratigraphy includes carbonate intervals, whereas comparable carbonate rocks are absent in the Central Sumatra Basin. This contrast can influence post-depositional cementation, dissolution, and the variability of log signatures. Under this design, this study evaluates whether depositional analogy alone supports model transferability and whether algorithmic adaptation is required to address basin-dependent geological controls. If significant statistical shifts are observed between these geologically similar formations, it confirms that secondary geological factors (such as diagenesis, compaction, or specific mineralogy) create a domain shift that necessitates the proposed unsupervised –Supervised Learning (UL-SL) framework.

For this analysis, four standard well log curves were utilized as input features for both basins: Gamma Ray (GR), Neutron Porosity (NPHI), Bulk Density (RHOB), and Photoelectric Factor (PEF). These logs were chosen for their combined sensitivity to lithology, mineralogy, and porosity, providing a comprehensive basis for identifying electrofacies and quantifying the statistical properties of the rock formations.

Evidence of statistical

A direct comparison of the data distributions between the source and target basins revealed significant disparities, contradicting the assumption that geological analogies imply statistical similarities. We analyzed the distribution of standardized well-log data to isolate the shape and spread of the data features from their absolute units.

As shown in Figure 6, the histograms exhibit clear domain shifts. Basin A displayed a broader, often bimodal distribution, particularly for gamma rays (GR), reflecting a distinct separation between sand and shale lithologies. In contrast, Basin B showed a significantly narrower and sharper distribution (higher kurtosis). This is most evident in

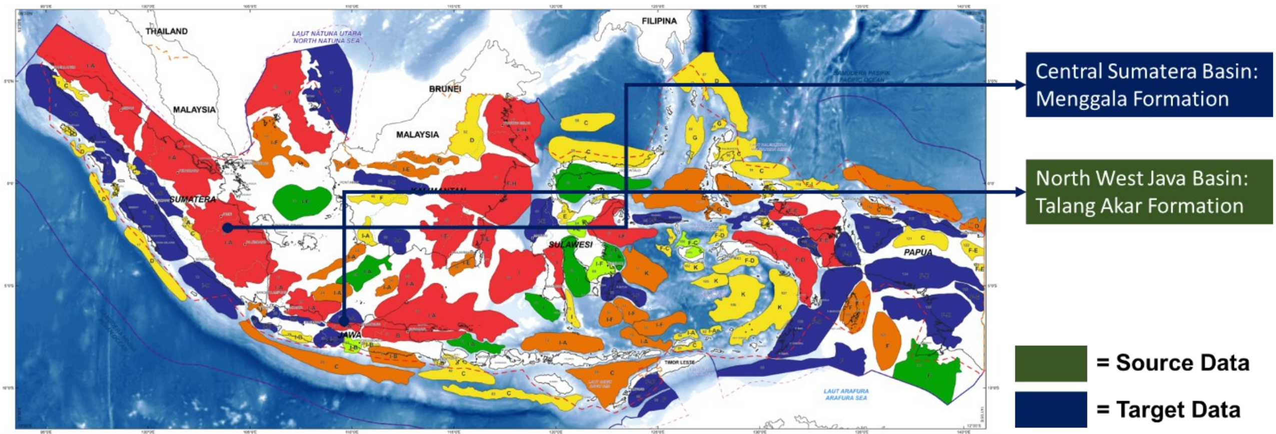

Figure 5. The basins location that shows the source and target data (Basin Map from Geological Agency, ESDM, 2022)

the bulk density (RHOB) log, where the Basin B data are concentrated in a tighter, higher-density range compared with the wider spread observed in Basin A. This shift implies a fundamental difference in rock compactness or lithological definition, despite shared fluvial depositional settings.

The pairplots show basin-dependent multivariate structure; therefore, direct transfer from the source basin to the target basin is not assured. As shown in Figures 7 and 8, the standard logs have different marginal distributions and pairwise relationships across basins when plotted in the original physical units. These differences appear in the spread, cluster overlap, and direction and strength of the observed trends.

In the source basin, the K-means clusters are more stratified with several feature combinations. In the target basin, the same clusters show stronger mixing and shifted separation boundaries, indicating different feature space geometries. These results suggest that a supervised model trained only on source basin data may learn interaction patterns that are not stable in the target basin, which can reduce cross-basin generalization.

Accordingly, cross-basin prediction requires an adaptation step, such as electrofacies mapping or feature space alignment, to obtain a representation that is more comparable across basins. This strong GR-PEF coupling in Basin B provides evidence of basin-specific heterogeneity. This suggests a distinct clay mineralogy in the Central Sumatra Basin, likely dominated by iron-rich clays (such as chlorite or illite) or associated heavy minerals that possess a higher atomic number, thereby increasing the PEF response. In contrast, the clays in Basin A did not exhibit this mineralogical signature.

This finding validates the argument presented in our literature review that second-order geological controls, such as provenance and mineralogy, create unique data structures. A machine learning model trained on Basin A would learn that "shale has low-to -moderate PEF," leading to substantial misclassification when applied to Basin B, where "shale exhibits high PEF." This empirically confirms that data distribution shifts are physical realities driven by geology and not merely statistical noise.

Quantification of cross-basin barriers

To test the validity of the industry assumption that models can be directly applied to analogous formations, we conducted a direct mapping experiment. In this scenario, the cluster centroids defined in source Basin A used to classify the electrofacies in target Basin B. If the geological analogy hypothesis is true, the centroids from Basin A should naturally overlap or reside near the native data structure of Basin B.

However, the experiment revealed significant divergence, as shown in Figure 9. The cluster centroids from the two basins did not overlap; instead, they separated. The average Euclidean distance between the mapped centroids and local centroids of Basin B was 21.81 in the standardized space. Given that standardized data typically fall within a range of ±3, a distance magnitude of 21.81 is not merely a statistical deviation but also physical evidence of a fundamental difference in rock properties. This confirms that the "center of gravity" for specific rock types has shifted drastically between the two basins owing to local geological factors despite sharing the same lithological names.

We quantified the severity of this domain shift using the adjusted Rand index (ARI) and normalized mutual information (NMI). The analysis yielded an ARI of 0.113, which is close to zero. This statistically refutes the hypothesis that geological analogies guarantee similar log responses. This low score indicates that the latent structures in the target basin were almost uncorrelated with the patterns learned from the source basin. The inability of the model to recognize these patterns is not an algorithmic failure but rather highlights the magnitude of the geological barrier that conventional studies fail to measure.

The practical consequences of this domain shift are shown in Figure 10. The side-by-side comparison between the "Pure" track (derived from native data) and the "Mapped" track (predicted using Basin A rules) shows a mismatch. Intervals identified locally as high-density shales were frequently misclassified as sandstones by the mapped model. This error occurs because the model failed to recognize the unique compaction signature of Basin B. These findings confirm that

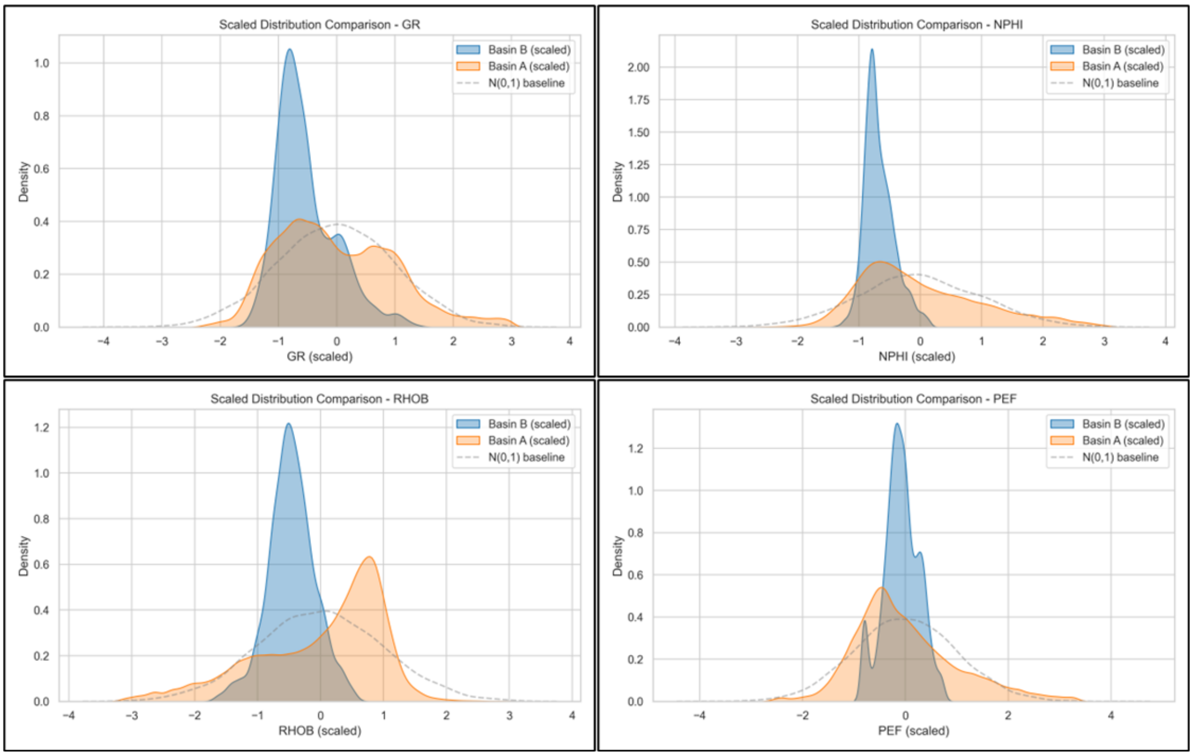

Figure 6. The Histogram show domain shift between basin

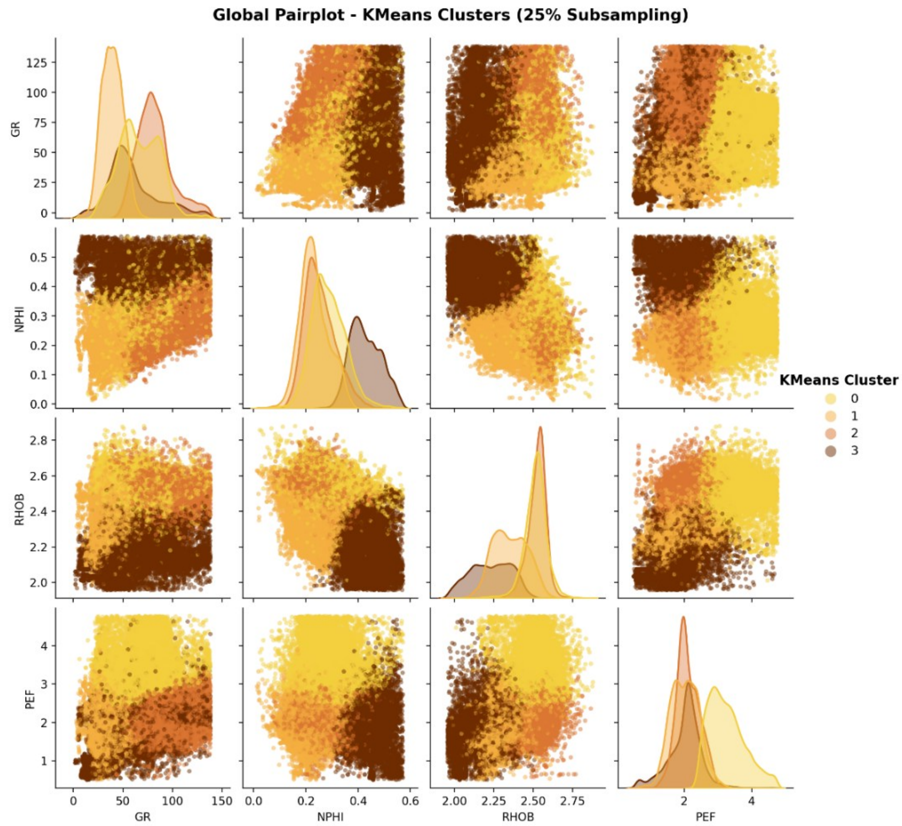

Figure 7. The pairplot for Basin A

applying models naively without domain adaptation leads to misleading reservoir interpretations.

DISCUSSION

The empirical findings presented in this study provide critical validation of the conceptual challenges hindering ML generalization. The significant statistical divergence between the Talang Akar and Menggala Formations, despite their shared classification as fluvial-deltaic reservoirs, offers fundamental insights that necessitate a paradigm shift in the definition of geological analogs for data-driven workflows.

Depositional vs. petrophysical

The central argument of this study is that geological and statistical similarities are not guaranteed. The industry currently relies on a depositional analog heuristic (Ismail et al., 2025), assuming that models trained on one fluvial system will naturally function on another. Our analysis proves this assumption to be flawed because of the stratification of geological controls. While the first-order control of the depositional environment establishes a shared sedimentary template, second-order controls, such as provenance, mineralogy, and diagenesis, dictate the final statistical distribution of well logs.

The most notable evidence of this is the correlation between Gamma Ray and Photoelectric Factor observed in Basin B which is noticeably absent in Basin A. In the source basin, the photoelectric factor is independent of gamma rays. However, in the target basin, the strong linear relationship indicates a distinct mineralogical signature. This specific relationship serves as a strong indicator of the unique mineralogical composition of the Central Sumatra Basin. Although direct mineralogical sampling is required for confirmation, the log response is characteristic of sediments containing elements with higher atomic numbers. This pattern is consistent with the presence of iron-rich clays or associated heavy minerals that increase the photoelectric response.

This empirical evidence directly validates the conceptual model presented in Figure 3. This diagram illustrates how diagenetic overprints and

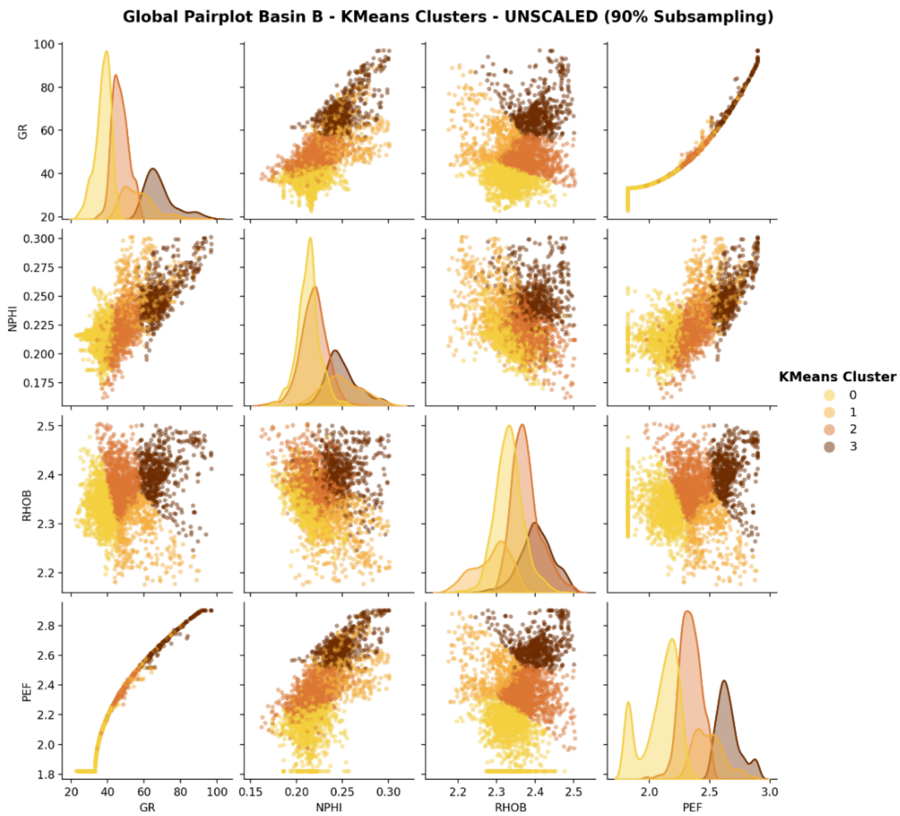

Figure 8. The pairplot for Basin B

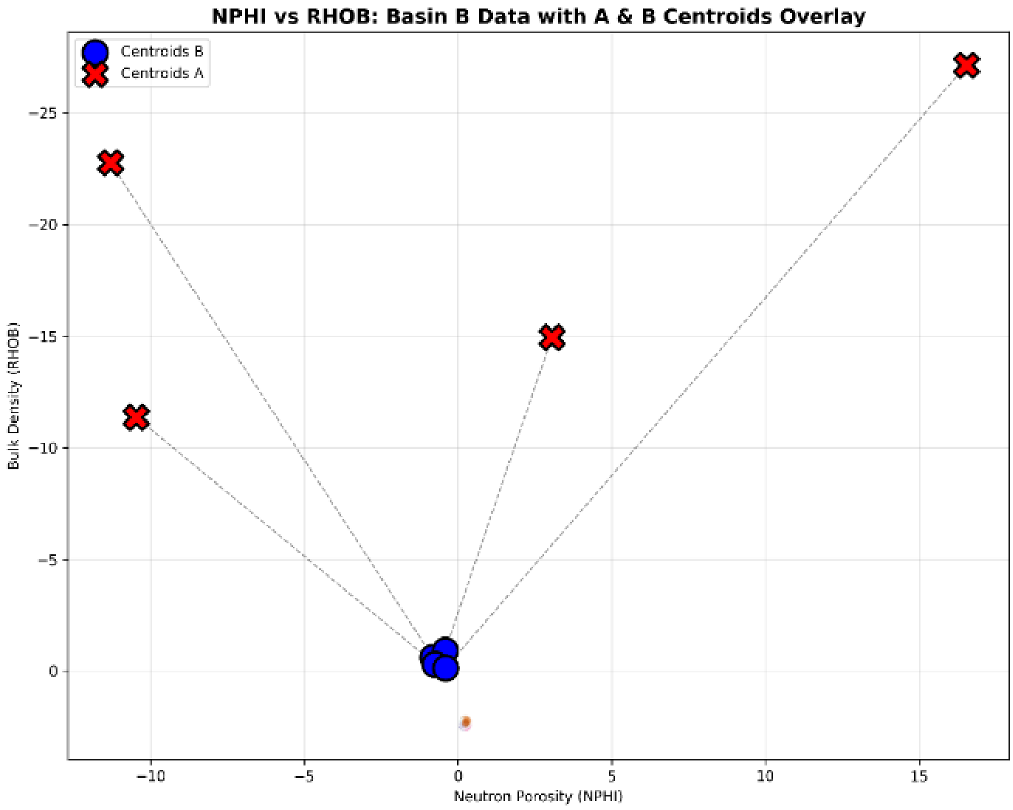

Figure 9. The plot showing centroids from the two basins do not overlap

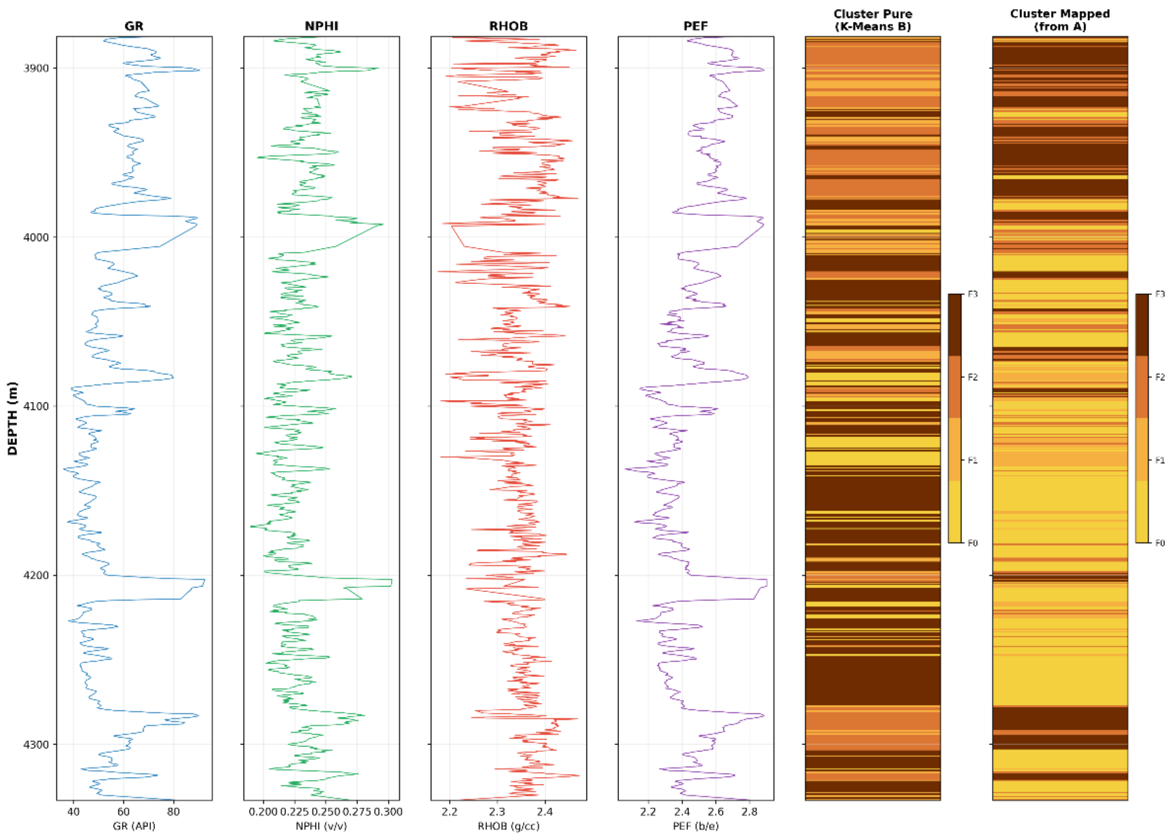

Figure 10. The log plot of X-1 well in Basin B

mineralogical changes can effectively decouple the final petrophysical properties from the initial depositional environment. Our results confirm this theoretical mechanism in real-world settings. The distinct density distribution and unique photoelectric response in Basin B reflect a specific compaction and diagenetic history that differs from that of Basin A, despite their shared sedimentary origin. Thus, although the two basins are depositional analogs, they are distinct petrophysical domains.

Considering these findings, we suggest that the concept of an analog in machine learning should be defined with greater precision and robustness. For a model to be directly transferable, the target basin must be a petrophysical analog rather than merely a depositional analog. A true petrophysical analog requires alignment of the depositional setting, mineralogical composition, and diagenetic history. The large Euclidean distance of 21.81 between the cluster centroids of basins A and B quantifies the gap between the two definitions. The low Adjusted Rand Index statistically confirms that without accounting for these second-order controls, the latent structures of the data remain incompatible.

Geological drivers of domain shift: ranked hypotheses and log-based proxies

The above results indicate that domain shift between Basin A and Basin B is not controlled by depositional analogy alone, but by second-order geological controls that alter mineralogical composition and rock-physics relationships expressed in well logs. Because direct mineralogical and diagenetic measurements are not uniformly available, the drivers below are presented as ranked hypotheses supported by log-based proxies, such as shifts in marginal distributions, changes in feature interactions, and differences in electrofacies cluster geometry.

- Mineralogical and lithological contrasts linked to basin stratigraphy

A first-order hypothesis is that basin-scale differences in lithology and mineralogy place the two datasets in different regions of multivariate log space, thereby altering cluster separability and cross-feature structure. Well-log cluster and electrofacies studies have shown that clusters can be calibrated to mineralogical variability and propagated through conventional logs, implying that mineralogical contrasts naturally translate into different cluster geometries across areas (Euzen & Power, 2012). In addition, standard log interpretation references highlight that the photoelectric effect (Pe/PEF) is strongly controlled by atomic number and is a robust lithology indicator with relatively low sensitivity to pore fluids, reinforcing that compositional differences can shift log behavior even when depositional labels are similar (Kansas Geological Survey, 2017). Mineral-controlled multivariate log trends have also been documented in log-based property prediction studies, which have shown that correlations among common logs depend on mineral content and should be treated as lithology-group dependent (Fuchs & Förster, 2014).

- Diagenetic overprint and compaction history

A second hypothesis is that differences in burial evolution, cementation, dissolution, and chemical compaction reorganize density–sonic–porosity relationships and change inter-log coupling, producing systematic domain shift even within broadly similar depositional systems. Studies of diagenetically altered mudstones demonstrate that normal compaction trends and log responses depend on temperature and stress history, indicating that basin-to-basin variations in diagenesis and compaction can invalidate a single transferable trend (Goulty & Sargent, 2016). Log-based studies also show depth-related changes in interval transit time associated with maturity and compaction effects, supporting that basin-dependent compaction histories can shift sonic-derived relationships (Lang, 1994). Together, these studies support interpreting cross-basin shifts in joint distributions among density, sonic, and porosity-sensitive logs as plausible indicators of different compaction–diagenesis histories (Goulty & Sargent, 2016; Lang, 1994).

- Shale fraction and clay fabric variability in fluvial–deltaic systems

A third hypothesis is that differences between basins in the shale fraction, clay mineral assemblage, and shale fabric modify the gamma response and its coupling with porosity- and density-sensitive logs. These effects can produce basin-specific feature interactions and electrofacies stratification. Spectral natural gamma ray studies in the Northwest Java Basin have shown that variations in K, Th, and U responses are linked to clay mineral content and shale-related characteristics. This finding supports the interpretation that subtle clay mineral changes can create systematic differences in log space, even within broadly similar depositional systems (Wibowo et al., 2020). Established workflows have also shown that gamma ray logs can be used to estimate clay content through optimized indices, reinforcing the shale fraction as a measurable control on log statistics and cluster structure (Diaz-Curiel et al., 2021). Independent validation studies comparing XRD-derived clay minerals with log-based cross-plot interpretations further indicate that clay type and clay content can alter multivariate log relationships and classification boundaries (Alaskari, 2018). In fine-grained intervals, clay effects on log responses are also critical for porosity evaluation, which is consistent with clay-related variability as a driver of domain shift in density, neutron, and porosity coupling (Zhu et al., 2023).

- Provenance and sediment texture effects

A fourth hypothesis is that provenance-driven texture and framework composition change porosity and permeability evolution, which then propagates to density–sonic–porosity distributions and electrofacies separability. Classic texture studies have shown that sorting and grain-size distribution exert strong control on porosity and permeability in unconsolidated sands, implying that provenance-related textural differences can create systematic differences in log-derived trends between basins (Beard & Weyl, 1973). A broader synthesis work on sandstone porosity evolution also identifies sorting, mineralogy, and maximum burial depth as dominant controls, aligning with the idea that cross-basin textural and burial differences shift petrophysical log behavior (Scherer, 1987).

- Secondary influences are not directly constrained by logs (poorly constrained by logs alone).

A fifth hypothesis is that climate-modulated weathering influences sediment composition indirectly, primarily through provenance and clay assemblages; however, it is difficult to isolate uniquely using conventional logs without supporting geochemical and mineralogical calibrations. Paleoclimate-oriented downhole logging literature frames key limitations in extracting quantitative climate signals from logs, including tool resolution and the complex origin of log responses, which are commonly confounded by lithology, porosity, and diagenesis (DeMenocal et al., 1992). Therefore, climate-related weathering is treated as a secondary driver in log-only domain shift interpretation unless external mineralogical and geochemical datasets are available (DeMenocal et al., 1992).

Overall, the ranked evidence supports interpreting mineralogical composition and post-depositional modification as dominant contributors to domain shift in this study. This study identified plausible drivers using log-based evidence; quantifying the relative dominance of each driver requires dedicated mineralogical, petrographic, and diagenetic datasets in future work.

Adaptive workflow strategy

The identification of this analog gap shifts the status of the proposed unsupervised–supervised learning framework from a theoretical option to a methodological necessity. In a traditional supervised learning approach, the model is rigid and forces new data to fit old rules. It lacks a mechanism to detect whether the definition of shale has changed owing to mineralogy. Unsupervised learning is critical in this regard.

By applying clustering algorithms to the target basin before prediction, the unsupervised component acted as a diagnostic scanner for mineralogy. It captures the intrinsic structure of the new data, such as the unique high-density clusters of Basin B, independent of external labels. These local clusters serve as a bridge, allowing the framework to map the statistical reality of the target basin to the semantic labels of the source basin. Therefore, this study argues that blind transfers are now obsolete. Robust cross-basin characterization requires an adaptive workflow in which unsupervised learning first quantifies the domain shift and then guides the supervised model to adapt its predictions. This geologically grounded approach is the only viable way to unlock the value of legacy data for frontier exploration.

CONCLUSION

The application of machine learning in petrophysics is currently hindered by widespread reliance on intra-basin validation. This common practice fails to assess how models perform in new geological settings and often yields optimistic metrics that do not hold in real-world exploration scenarios. This study addresses this critical gap by providing quantitative empirical evidence regarding the limits of current model generalization assumptions.

A comparative analysis of the Talang Akar and Menggala formations revealed that sharing an analogous fluvial-deltaic depositional environment does not guarantee statistical alignment. We identified a significant domain shift, which was quantified by a critically low adjusted Rand index of 0.113. This statistical divergence confirms that secondary geological controls, such as distinct clay mineralogy and differing compaction regimes, create unique data signatures for each basin, despite their shared sedimentary origins.

These findings provide concrete proof that the direct transfer of models between analog basins carries a high risk of reservoir mischaracterization without algorithmic intervention. These findings do not represent a research barrier but rather serve as a strong justification for the subsequent phases of this study. Therefore, future work should prioritize two strategic directions: data enrichment and methodological advancement. We plan to expand the training datasets in Basin A and the target datasets in Basin B to determine whether the observed domain shift can be mitigated by simply increasing the sample size and variety. Furthermore, we intend to introduce a third dataset from Basin C in the South Sumatra Basin to serve as a comparative benchmark. This control group will help verify whether centroid separation is a unique anomaly between the current basins or a consistent universal challenge in cross-basin validation.

Finally, the research will proceed to the full implementation of domain adaptation techniques and the proposed integrated unsupervised–supervised learning model. The next phase aims to actively bridge the identified feature gap by utilizing local cluster labels as guiding inputs. This strategy ensures that the model adapts to the specific mineralogical and diagenetic realities of the target basin, rather than imposing rigid assumptions from the training data. By validating the need for domain adaptation, this study lays the necessary groundwork for developing a robust and geologically consistent reservoir characterization workflow for frontier areas.

GLOSSARY OF TERMS AND SYMBOLS

Terms & Symbols	Definition	Unit
ARI	Adjusted Rand index, a clustering evaluation metric corrected for chance agreement	-
CNN	Convolutional neural networks are deep-learning architectures designed for spatial or sequential feature extraction.	-
DA	Domain adaptation techniques are used to align a model across different data distributions.	-
DL	Deep Learning; machine learning methods utilizing multi-layered neural networks.	-
GR	Gamma Ray log	API
ML	Machine Learning; algorithms that identify patterns in data for prediction or classification.	-
NMI	Normalized mutual information, an information-theoretic metric for evaluating clustering performance.	-
NN	neural network, a computational model based on interconnected	-

	artificial neurons.	-
NPHI	Neutron porosity log, a formation measurement based on the hydrogen index to estimate porosity.	v/v
PEF	Photoelectric Factor log	b/e
RHOB	Bulk Density log	g/cm^3
SL	Supervised learning is a modeling approach that utilizes labeled input-output data for training.	-
TL	Transfer learning: applying knowledge from a pre-trained model to a new, related task.	-
UL	unsupervised learning algorithms that discover hidden structures within unlabeled data.	-
UL-SL	Unsupervised–Supervised Learning: a hybrid framework that combines unlabeled data exploration with supervised training.	-

ACKNOWLEDGEMENT

The authors wish to acknowledge LAPI ITB for supporting this study. We also express our gratitude to the Doctoral Program in Petroleum Engineering, Institut Teknologi Bandung (ITB), and the Petroleum Engineering Study Program, Tanri Abeng University, for the academic support and facilities that made this research possible.

AUTHOR CONTRIBUTION

R.C. Rohmana: conceptualization, methodology, investigation, data curation, formal analysis, software, visualization, writing original draft, writing – review & editing. T. Ariadji: Supervision (primary), Conceptualization, Methodology, Writing – Review & Editing, Validation. A. Yasutra: Supervision (co), Methodology, Writing – Review & Editing, Validation. D. Irawan: Supervision (co), Methodology, Writing – Review & Editing, Validation.

REFERENCES

Abdel-Fattah, M. I., (2015), Impact of depositional environment on petrophysical reservoir characteristics in Obaiyed Field, Western Desert, Egypt. Arabian Journal of Geosciences, 8(11), 9301–9314. https://doi.org/10.1007/s12517-015-1913-5.

Alaskari, G. M. K., (2018), XRD evaluation of clay minerals in shaley formation and its comparison with cross plotting of log data. Progress in Petrochemical Science, 1(3), 64–67. https://doi.org/10.31031/PPS.2018.01.000513.

Badan Geologi, Kementerian ESDM, (2022), Peta cekungan sedimen Indonesia 2022. https://www.esdm.go.id/assets/media/content/content-peta-cekungan-sedimen-indonesia-2022.pdf.

Beard, D. C., & Weyl, P. K., (1973), Influence of texture on porosity and permeability of unconsolidated sand. AAPG bulletin, 57(2), 349-369.

Brackenridge, R. E., Demyanov, V., Vashutin, O., & Nigmatullin, R. (2022). Improving subsurface characterisation with 'big data' mining and machine learning. Energies, 15(3), 1070. https://doi.org/10.3390/en15031070.

Candra, A. D., Rahalintar, P., Sulistiyono, S., & Prabowo, U. N., (2024), Comparison of facies estimation of well log data using machine learning. Scientific Contributions Oil and Gas, 47(1), 21–30. https://doi.org/10.29017/SCOG.47.1.1593.

Cuddy, S. J., (2000), Litho-facies and permeability prediction from electrical logs using fuzzy logic. Society of Petroleum Engineers.

DeMenocal, P. B., Bristow, J. F., & Stein, R., (1992), Paleoclimatic applications of downhole logs: Pliocene-Pleistocene results from Hole 798B, Sea of Japan. In Proceedings of the Ocean Drilling Program Vol. 127, No. Pt. 1, p. 337). Ocean Drilling Program.

Díaz-Curiel, J., Miguel, M. J., Biosca, B., & Arévalo-Lomas, L., (2021), Gamma ray log to estimate clay content in the layers of water boreholes. Journal of Applied Geophysics, 195, 104481.

Dong, Y., Zhang, Y., Liu, F., & Cheng, X., (2021), Reservoir production prediction model based on a stacked LSTM network and transfer learning. ACS Omega, 6(50), 34700–34711. https://doi.org/10.1021/acsomega.1c05132.

Dramsch, J. S., (2020), 70 years of machine learning in geoscience in review. Advances in Geophysics, 61, 1–55. Elsevier. https://doi.org/10.1016/bs.agph.2020.08.002.

Dwihusna, N., (2020), Seismic and Well Log Based Machine Learning Facies Classification in the Panoma-Hugoton Field, Kansas and Raudhatain Field, North Kuwait. Colorado School of Mines.

Euzen, T., & Power, M. R., (2012), Well log cluster analysis and electrofacies classification: a probabilistic approach for integrating log with mineralogical data. In 2012 CSPG CSEG CWLS Convention.

Fuchs, S., & Förster, A., (2014), Well-log based prediction of thermal conductivity of sedimentary successions: a case study from the North German Basin. Geophysical Journal International, 196(1), 291-311.

Gonçalves, C. A., Harvey, P. K., & Lovell, M. A., (1997), Prediction of petrophysical parameter logs using a multilayer backpropagation neural network. Geological Society, London, Special Publications, 122(1), 169–180. https://doi.org/10.1144/GSL.SP.1997.122.01.13.

Goulty, N. R., & Sargent, C., (2016), Compaction of diagenetically altered mudstones–Part 2: Implications for pore pressure estimation. Marine and Petroleum Geology, 77, 806-818.

Gu, Y., Bao, Z., Song, X., Patil, S., & Ling, K., (2019), Complex lithology prediction using probabilistic neural network improved by continuous restricted Boltzmann machine and particle swarm optimization. Journal of Petroleum Science and Engineering, 179, 966–978. https://doi.org/10.1016/j.petrol.2019.05.032.

He, M., Gu, H., & Wan, H., (2020), Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir. Journal of Petroleum Science and Engineering, 194, 107498. https://doi.org/10.1016/j.petrol.2020.107498.

Imamverdiyev, Y., & Sukhostat, L., (2019), Lithological facies classification using deep convolutional neural network. Journal of Petroleum Science and Engineering, 174, 216–228. https://doi.org/10.1016/j.petrol.2018.11.023.

Iraji, S., Soltanmohammadi, R., Matheus, G. F., Basso, M., & Vidal, A. C., (2023), Application of unsupervised learning and deep learning for rock type prediction and petrophysical characterization using multi-scale data. Geoenergy Science and Engineering, 230, 212241. https://doi.org/10.1016/j.geoen.2023.212241.

Ismail, H., Shehata, A. A., Ismail, A., & Attia, T. E., (2025), Facies analysis and depositional environments interpretation and petrophysical evaluation of the Pliocene Kafr El-Sheikh reservoirs at Sapphire Field, West Delta Deep Marine, Egypt. Alfarama Journal of Basic & Applied Sciences, 6(2), 184–198. https://doi.org/10.21608/ajbas.2025.361285.1250.

Jafari, J., Mahboubi, A., Moussavi-Harami, R., & Al-Aasm, I. S., (2020), The effects of diagenesis on the petrophysical and geochemical attributes of the Asmari Formation, Marun oil field, southwest Iran. Petroleum Science, 17(2), 292–316. https://doi.org/10.1007/s12182-019-00421-0.

Kansas Geological Survey, (2017), The photoelectric factor (PeF). Retrieved February 26, 2026, from https://www.kgs.ku.edu/Publications/Bulletins/LA/06_photo.html.

Khan, H., Srivastav, A., & Mishra, A. K., (2019), Estimation of permeability of a reservoir using deep learning algorithms on well logs. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3349570.

Kompantsev, G., (2024), Transfer learning for variable well locations and permeability distributions in physics-aware deep learning reservoir simulation proxy models. In SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/223505-STU.

Lang, W. H., (1994), Compaction/diagenesis of sediments and compaction gradients in relation to interval transit time. The log analyst, 35(04).

Lee, S. H., & Datta-Gupta, A., (1999), Electrofacies characterization and permeability predictions in carbonate reservoirs: Role of multivariate analysis and nonparametric regression. Society of Petroleum Engineers.

Leila, M., & Moscariello, A., (2018), Depositional and petrophysical controls on the volumes of hydrocarbons trapped in the Messinian reservoirs, onshore Nile Delta, Egypt. Petroleum, 4(3), 250–267. https://doi.org/10.1016/j.petlm.2018.04.003.

Li, Z., Kang, Y., Feng, D., Wang, X.-M., Lv, W., Chang, J., & Zheng, W. X., (2020), Semi-supervised learning for lithology identification using Laplacian support vector machine. Journal of Petroleum Science and Engineering, 195, 107510. https://doi.org/10.1016/j.petrol.2020.107510.

Li, Z., Wang, Z., Wei, Z., Zhou, X., Wang, Y., Huai, B., Liu, Q., Yuan, N. J., Gong, R., & Chen, E., (2021), Cross-oilfield reservoir classification via multi-scale sensor knowledge transfer. Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), 4215–4223. https://doi.org/10.1609/aaai.v35i5.16545.

Lima, M. C. O., Pontedeiro, E. M., Ramirez, M. G., Favoreto, J., Dos Santos, H. N., Van Genuchten, M. Th., Borghi, L., Couto, P., & Raoof, A., (2022), Impacts of mineralogy on petrophysical properties. Transport in Porous Media, 145(1), 103–125. https://doi.org/10.1007/s11242-022-01829-w.

Lorenz, J. C., Sattler, A. A., & Stein, C. L., (1989), The effects of depositional environment on petrophysical properties of Mesaverde reservoirs, Northwestern Colorado. Society of Petroleum Engineers.

Magoba, M., Opuwari, M., & Liu, K., (2024), The effect of diagenetic minerals on the petrophysical properties of sandstone reservoir: A case study of the Upper Shallow Marine sandstones in the Central Bredasdorp Basin, offshore South Africa. Minerals, 14(4), 396. https://doi.org/10.3390/min14040396.

Merletti, G. D., Spain, D. R., Melick, J., Armitage, P., Hamman, J., Shabro, V., & Gramin, P., (2017), Integration of depositional, petrophysical, and petrographic facies for predicting permeability in tight gas reservoirs. Interpretation, 5(2), SE29–SE41. https://doi.org/10.1190/INT-2016-0112.1.

Misra, S., Elkady, M., Kumar, V., Odi, U., & Silver, A., (2024), Use of transfer learning in shale production forecasting. In International Petroleum Technology Conference. https://doi.org/10.2523/IPTC-23438-MS.

Mohaghegh, S., Arefi, R., Bilgesu, I., Ameri, S., & Rose, D., (1995), Design and development of an artificial neural network for estimation of formation permeability. SPE Computer Applications, 7(06), 151–154. https://doi.org/10.2118/28237-PA.

Mohaghegh, S., Bogdan, B., & Ameri, S., (1997), Permeability determination from well log data. SPE Formation Evaluation, 12(3), 170–174.

Nugroho, I. D. R., Trisna, M. D., & Saroji, S., (2024), An implementation of XGBoost and Random Forest algorithm to estimate effective porosity and permeability on well log data at Fajar Field, South Sumatera Basin, Indonesia. Indonesian Journal of Applied Physics, 14(2), 271. https://doi.org/10.13057/ijap.v14i2.82901.

Omoboriowo, A. O., Chiadikobi, K. C., & Chiaghanam, O. I., (2012), Depositional environment and petrophysical characteristics of “LEPA” reservoir, Amma Field, Eastern Niger Delta, Nigeria. International Journal of Pure and Applied Sciences and Technology, 10(2), 38–61.

Pan, W., (2022), Reservoir Description via Statistical and Machine-Learning Approaches. The University of Texas at Austin.

Pratama, H., (2018), Machine learning: Using optimized KNN (k-nearest neighbors) to predict the facies classifications. In Proceedings of the 13th SEGJ International Symposium (pp. 538–541). https://doi.org/10.1190/SEGJ2018-139.1.

Rashid, F., Hussein, D., Glover, P. W. J., Lorinczi, P., & Lawrence, J. A., (2022), Quantitative diagenesis: Methods for studying the evolution of the physical properties of tight carbonate reservoir rocks. Marine and Petroleum Geology, 139, 105603. https://doi.org/10.1016/j.marpetgeo.2022.105603.

Sarhan, M. A., (2022), Geophysical appraisal of the Abu Madi gas reservoir, Nile Delta Basin, Egypt: Implications for the tectonic effect on the lateral distribution of petrophysical parameters. Petroleum Research, 7(4), 511–520. https://doi.org/10.1016/j.ptlrs.2022.03.002.

Scherer, M., (1987), Parameters influencing porosity in sandstones: a model for sandstone porosity prediction. AAPG bulletin, 71(5), 485-491.

Serra, O., (1984), Fundamentals of well-log interpretation: The acquisition of logging data. Elsevier.

Singh, H., Seol, Y., & Myshakin, E. M., (2020), Automated well-log processing and lithology classification by identifying optimal features through unsupervised and supervised machine-learning algorithms. SPE Journal, 25(05), 2778–2800. https://doi.org/10.2118/202477-PA.

Smeraglia, L., Trippetta, F., Carminati, E., & Mollo, S., (2014), Tectonic control on the petrophysical properties of foredeep sandstone in the Central Apennines, Italy. Journal of Geophysical Research: Solid Earth, 119(12), 9077–9094. https://doi.org/10.1002/2014JB011221.

Talebkeikhah, M., Sadeghtabaghi, Z., & Shabani, M., (2021), A comparison of machine learning approaches for prediction of permeability using well log data in the hydrocarbon reservoirs. Journal of Human, Earth, and Future, 2(2), 82–99. https://doi.org/10.28991/HEF-2021-02-02-01.

Ulil, M. R., Winardhi, S., & Dinanto, E., (2025), Machine learning-based prediction of shear wave velocity: Performance evaluation of Bi-GRU, ANN, and the Greenberg-Castagna empirical method. Scientific Contributions Oil and Gas, 48(3), 125–134. https://doi.org/10.29017/scog.v48i3.1797.

Verma, S., Bhattacharya, S., Chowdhury, N. U. M. K., & Tian, M., (2021), A new workflow for multi-well lithofacies interpretation integrating joint petrophysical inversion, unsupervised, and supervised machine learning. In First International Meeting for Applied Geoscience & Energy (pp. 2213–2217). https://doi.org/10.1190/segam2021-3584118.1.

Wibowo, R. C., Pertiwi, A. P., & Kurniati, S., (2020), Identification of Clay Mineral Content Using Spectral Gamma Ray on Y1 Well in Karawang Area, West Java, Indonesia. Journal of geoscience, engineering, environment, and technology, 5(3), 136-142.

Wong, P. M., Gedeon, T. D., & Taggart, I. J., (1995), An improved technique in porosity prediction: A neural network approach. IEEE Transactions on Geoscience and Remote Sensing, 33(4), 971–980. https://doi.org/10.1109/36.406683.

Wood, D. A., (2020), Predicting porosity, permeability and water saturation applying an optimized nearest-neighbour, machine-learning and data-mining network of well-log data. Journal of Petroleum Science and Engineering, 184, 106587. https://doi.org/10.1016/j.petrol.2019.106587.

Yang, L., Wang, S., Chen, X., Chen, W., Saad, O. M., Zhou, X., Pham, N., Geng, Z., Fomel, S., & Chen, Y., (2023), High-fidelity permeability and porosity prediction using deep learning with the self-attention mechanism. IEEE Transactions on Neural Networks and Learning Systems, 34(7), 3429–3443. https://doi.org/10.1109/TNNLS.2022.3157765.

Yao, J., Liu, Q., Liu, W., Liu, Y., Chen, X., & Pan, M., (2020), 3D reservoir geological modeling algorithm based on a deep feedforward neural network: A case study of the Delta Reservoir of Upper Urho Formation in the X Area of Karamay, Xinjiang, China. Energies, 13(24). https://doi.org/10.3390/en13246699.

Zainuri, A. P. P., Sinurat, P. D., Irawan, D., & Sasongko, H., (2023), Trap prevention in machine learning in prediction of petrophysical parameters: A case study in The Field X. Scientific Contributions Oil and Gas, 46(3), 115–127. https://doi.org/10.29017/SCOG.46.3.1586.

Zhang, Y., Hu, J., & Zhang, Q., (2021), Application of locality preserving projection-based unsupervised learning in predicting the oil production for low-permeability reservoirs. SPE Journal, 26(03), 1302–1313. https://doi.org/10.2118/201231-PA.

Zhu, L. Q., Sun, J., Zhou, X. Q., Li, Q. P., Fan, Q., Wu, S. L., & Wu, S. G., (2023), Well logging evaluation of fine-grained hydrate-bearing sediment reservoirs: Considering the effect of clay content. Petroleum Science, 20(2), 879-892.