Journal of Education for Pure Science- University of Thi-Qar
Vol.16, No. 02 (2026)
DOI: https://doi.org/10.32792/jeps.v16i2.758

Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

Persistent Local Topological Indicators of Feature Relevance in Subsampled Reservoir Logs

Sajaad Kadhum Hassan1[1], and Hana' Murtadha Ali[*,2]

[1] Department of Mathematics, College of Science, University of Basrah, Basra, 61004, Iraq
[2] Department of Mathematics, College of Science, University of Basrah, Basra, 61004, Iraq

[*] Corresponding email: hana.ali@uobasrah.edu.iq

Received 25/ 07 /2025, Accepted 25 / 08/2026, Published 01/ 06 /2026

Abstract:

This study evaluates the local stability of topological complexity as a model-agnostic proxy for feature importance in subsurface machine learning. Building on our previous work linking Betti-0 persistence metrics to predictive utility, we apply the framework to eight distinct well configurations, ranging from full-reservoir integration to single wells, to assess whether topological simplicity reliably indicates predictive relevance at localized scales. For each configuration, Betti-0 lifetimes were computed across all valid log subsets derived from seven geophysical features. Through additive marginalization, log-specific complexity scores were generated and compared with feature importance rankings from Random Forest models. A consistent inverse relationship emerged: logs with lower topological complexity (low β_0^{sum}) tended to rank higher in predictive relevance. In all cases, Spearman correlations exceeded (-0.85) and were statistically significant, confirming the strength of this alignment. Notably, well (15/9-F-1 A) recorded the lowest correlation ($\rho = -0.85$) among individual wells due to the vertical imbalance in landmark coverage. However, its alignment improved to a perfect inverse ($\rho = -1.0$) when paired with a geologically stable neighbor, highlighting the method's robustness to local heterogeneity. These findings confirm that persistent homology captures meaningful geometric structure in well log data and provides a scalable, interpretable, and architecture-independent approach to feature selection. The proposed framework remains effective under data constraints and geological variability, supporting its integration into real-time log prioritization and geoscientific modeling workflows.

Keywords: Betti number, Feature selection, Geophysical logs, Persistent homology, Sonic transit time, Topological data analysis.

1-Introduction Section

In petroleum reservoir modeling, applying machine learning (ML) to well-log data presents substantial challenges due to the complexity and volume of the data. McDonald (2021) highlights that modern well-log datasets exhibit the full spectrum of Big Data characteristics: volume, variety, velocity, veracity, and value, and require rigorous quality control before any reliable predictive modeling can be performed [1]. These challenges are further amplified when incorporating topological data analysis (TDA), a mathematically rich approach that introduces high computational and memory demands. However, such complexity becomes justifiable in the context of rising digitalization needs, limited access to in-situ data, and cost-cutting strategies in a low-price oil economy. One critical target for prediction is the compressional sonic transit time (DT), a key petrophysical parameter used to infer rock

elasticity and formation characteristics. However, DT readings are frequently incomplete or missing across many wells. To address this, researchers increasingly rely on predictive models that utilize other commonly recorded logs, such as gamma ray (GR), bulk density (RHOB), neutron porosity (NPHI), and resistivity (RT), to estimate DT with minimal error [2], [3].

While conventional machine learning models, such as random forests or neural networks [3], [4], can achieve good predictive performance, their feature selection processes remain inherently model-dependent. Impurity-based and permutation-based importance scores, for instance, reflect how each log influences a given model's internal decision-making process, but they offer little insight into the intrinsic structure of the input data. Traditional feature-important methods focus on how models interpret data, not on what the data inherently represents. In contrast, our approach leverages persistent homology to quantify the geometric saturation of well-log data in depth, offering a model-agnostic perspective rooted in the data's topological configuration. This allows us to uncover structural patterns that may stem from geological variability or even acquisition-related inconsistencies that are often invisible to traditional model-centric evaluations. By shifting the focus from predictive behavior to structural interpretation, this topological perspective not only enhances the transparency of feature assessment but also provides a foundation for evaluating well-log informativeness across varying data conditions.

In our previous study [5], we addressed this gap by introducing a model-independent framework based on persistent homology, a core tool in TDA. We proposed four interpretable descriptors derived from the lifetimes of connected components (Betti-0 features), enabling a topological measure of log complexity. These descriptors were then linearly marginalized across multiple log combinations to infer each log's contribution to the prediction process. In this study, we seek to validate the robustness and generalizability of this framework by applying it to individual wells and small-scale well groupings, rather than full-reservoir data. This localized analysis allows us to assess whether the inverse relationship between topological complexity and feature importance observed previously persists when working with sparse, site-specific data. Our findings affirm the presence of stable predictive signals embedded in the topological structure of geophysical logs, regardless of spatial scope or data completeness.

This research enhances our previous topological framework by confirming its consistency and dependability at a localized scale, especially for individual wells and small well clusters. We investigate the persistence of the negative connection between topological complexity and predictive relevance, previously shown at the full-reservoir level, under site-specific data restrictions. We calculate four interpretable Betti-0 lifespan descriptors (sum, mean, variance, and maximum) for each input log to achieve this goal. To figure out how much they each contribute to the overall topology, a linear marginalization method is used. After that, these complexity metrics are compared to feature significance ratings from random forest models that were trained separately on single wells or pairs of wells. Our results show that topological simplicity is always a strong and reliable sign of how informative a log is. The findings further show that our persistent homology-based approach is not only scalable and model-agnostic, but also strong against subsampling and local variability. This makes it even more useful in real-world geophysical situations when data is sparse or fragmented.

2- Related Work Section

The prediction of compressional sonic transit time (DT) using alternative well logs has been the focus of several recent studies. For instance, Nero et al. (2023) applied ensemble models such as Random Forest and XGBoost to capture nonlinear relationships among logs, achieving high predictive accuracy [3]. Similarly, Wang et al. (2023) employed NGBoost to reconstruct missing DT and DTS logs, incorporating SHAP-based interpretability to evaluate feature contributions and interactions [2]. These efforts highlight the ongoing need to determine which logs are most informative for DT prediction. As a result, a variety of feature importance methods have been explored, including impurity-based measures [6], permutation-based sensitivity analyses [7], and model-agnostic frameworks such as SHAP [8], [9]. However, all these approaches are inherently model-dependent and thus may not reflect the intrinsic geometric or structural properties of the input data.

To address this limitation, Topological Data Analysis (TDA) has emerged as a powerful, model-independent alternative. Situated at the intersection of topology, geometry, and data science, TDA aims to extract qualitative insights from high-dimensional datasets using tools from algebraic topology. Unlike traditional statistical methods, which typically assume linearity or rely on local metrics, TDA focuses on the shape of data ,its connectivity, loops, and voids, as a means of uncovering latent structures that may be obscured by noise or dimensionality [10].

At the heart of TDA lies the idea that even unstructured data, such as a point cloud $X = \{x_i\}_{i\in I}$ in $\mathbb{R}^n$, possesses an underlying topological structure. This structure can be inferred through the construction of simplicial complexes, such as the Vietoris–Rips, Čech, or Alpha complexes, which encode how data points are connected under a chosen

proximity scale[10]. Once a simplicial complex K is constructed, algebraic techniques like homology $H_k(K)$, $k \geq 0$, are used to quantify topological features across dimensions, H_0 captures connected components, H_1 loops, and H_2 voids or cavities. These features are provably stable under small perturbations, making TDA highly robust to noise and variation in data[11].

A major strength of TDA lies in its model-independence and its capacity to operate across multiple scales. Through persistent homology, TDA tracks the evolution (birth and death) of topological features as the scale parameter varies, distinguishing meaningful structure from topological noise. This methodology has been successfully applied in a wide range of fields, including genomics, neuroscience, material science, and signal processing, demonstrating its ability to identify clusters, characterize shapes, and quantify structural complexity where traditional methods fall short [12], [13], [14].

However, persistent homology, particularly when built using Vietoris–Rips filtrations, faces scalability challenges, as computational complexity grows exponentially with dataset size [14], [15]. To mitigate this, we previously introduced a depth-wise downsampling strategy that preserves the stratigraphic structure of well-log data while reducing computational cost. This approach aligns with recent advances and is essential for making TDA tractable in geophysical applications.

3- Methodology Section

We will first examine the fundamental principles of persistent homology as used in Topological Data Analysis (TDA), followed by an overview of the geological context of the Volve Field reservoir. Both viewpoints are crucial for comprehending the design and interpretation of our methodology.

3.1. Homology: [16]

Let X be a topological space. By the q^{th} homology group of X we mean $H_q(X) = \mathbb{Z}^{\beta_q}$, where β_q is the number of the q-dimension holes of X. That is, β_0 forms the number of separate connected components (clusters); β_1 represents circle; β_2 are the gaps, and β_q is called q^{th} Betti number of X. In general, the Homology $H(X)$ is defined to be $\{H_q(X)\}_{q=0}^{\infty}$.

3.2. Vietoris-Rips Filtration [10]

The selection of the filtration determines how to create a process of simplicial complexes across a dataset at different scales, and therefore the generation of persistent homology for a certain point cloud $X = \{x_i\}_{i\in I} \subseteq \mathbb{R}^n$. The underlying topological structure may be better understood by using various filtrations. This research will make use of the Vietoris-Rips Complex and its construction approach in the following method:

1. Let $X \subseteq \mathbb{R}^n$ be a finite point cloud $X \subseteq \mathbb{R}^n$ and $\varepsilon > 0$. The Vietoris-Rips complex $VR_\varepsilon(X)$ contains all simplexes that its set of vertices is a $\sigma \subseteq X$ that satisfied $diam(\sigma) = Max_{x,y\in\sigma}\{d(x,y)\}$.
2. Let $\varepsilon_{inf} > 0$ be a real number with its Vietoris-Rips complex $VR_{\varepsilon_{inf}}(X)$ has the property for every $\varepsilon > \varepsilon_{inf}$, $VR_\varepsilon(X) = VR_{\varepsilon_{inf}}(X)$.
3. Let, $f: VR_{\varepsilon_{inf}}(X) \to \mathbb{R}$ be a real valued function defined as: $f(\sigma) \leq f(\rho)$ if and only if $\sigma \subseteq \rho$, for all $\sigma, \rho \in VR_{\varepsilon_{inf}}(X)$.
4. Assume $-\infty = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_d = \varepsilon_{inf}$ be the functions values of the simplexes of simplices of $VR_{\varepsilon_{inf}}(X)$ that satisfied if $0 \leq i < d$ and $\varepsilon_i < \varepsilon < \varepsilon_{i+1}$, then $VR_\varepsilon(X) = VR_{\varepsilon_i}(X)$.
5. For $0 \leq i < d$, put $VR_i = f^{-1}(-\infty, \varepsilon_i)$. Construct the $VR_{\varepsilon_{inf}}(X)$-filtration to be:

$$\emptyset = VR_0 \subseteq VR_1 \subseteq \ldots \subseteq VR_d = VR_{\varepsilon_{inf}}(X) \ldots (1)$$

3.3. Persistent Homology [10]

Assume we have the assumption of the item (3.2.). The persistent homology of X can be constructed as follows:

1. For $0 \leq i \leq d$, the simplicial complex VR_i defined a free Abelian group that generated by all its q-simplices denoted by $C_q(VR_i)$, $0 \leq q \leq \dim(VR_i) = \delta_i$. The relation between any simplex and its face defines the chain complex:

$$0 \longrightarrow C_{\delta_i}(VR_i) \xrightarrow{\partial^i_{\delta_i}} C_{\delta_i-1}(VR_i) \xrightarrow{\partial^i_{\delta_i-1}} \cdots \xrightarrow{\partial^i_2} C_1(VR_i) \xrightarrow{\partial^i_1} C_0(VR_i) \xrightarrow{\partial^i_0} 0 \ \ldots (2)$$

3. Thus, for $0 \leq i \leq j \leq d$, let $f^{i,j}: VR_i \to VR_j$ be the inclusion map. if $\delta_i < \delta_j$, then we have an inclusion chain map $f^{i,j}: C_*(VR_i) \to C_*(VR_j)$. For each dimension q, if $H_q(VR_i)$ and $H_q(VR_j)$ be the q^{th} homological

group for VR_i and VR_j respectively, and $f_q^{i,j}: H_q(VR_i) \to H_q(VR_j)$ be the induced homomorphism between them, then the $q^{th}-$persistent homology groups $H_q^{i,j}(X)$ are defined to be, $H_q^{i,j}(X) = Im(f_q^{i,j})$, for $0 \leq i < j \leq d$. That is, $H_q^{i,j}(X)$ consist of the homology classes of K_i that are still alive at K_j.

5. A persistence barcode is a visual representation of the lifespan of the homology classes that represent the topological features, such as connected components, loops, or voids, across a range of filtration scales. Each feature is depicted as a horizontal bar; the left endpoint marks its *birth* (when it first appears), and the right endpoint marks its *death* (when it disappears or merges with another feature). Longer bars indicate more persistent, and typically more meaningful, structures in the data, while shorter bars are often considered noise.

Figure 1 shows the topological development and persistence barcode of an input dataset. The point cloud changes shape slowly as the Vietoris–Rips filtering parameter ε travels from left to right. When $\varepsilon = 0.02$, each data point makes its own linked component, which gives a large Betti-0 count $\beta_0 = 38$, but there are no higher-order features (loops) ($\beta_1 = 0$). At $\varepsilon = 0.25$, components begin to merge, reducing the number of clusters ($\beta_0 = 30$). At $\varepsilon = 0.50$, only four clusters remain ($\beta_0 = 4$), and three short-lived loops emerge ($\beta_1 = 3$). Finally, at $\varepsilon = 0.75$, all points are connected into a single component ($\beta_0 = 1$) and loops vanish ($\beta_1 = 0$). The persistence barcode (bottom panel) consistently reflects these dynamics, with red bars representing connected components (β_0) and green bars representing loops (β_1). This consistency highlights the principle of persistent homology: features are born as ε increases, persist for a range of scales, and eventually die as structures merge, forming the descriptors used in our feature selection framework.

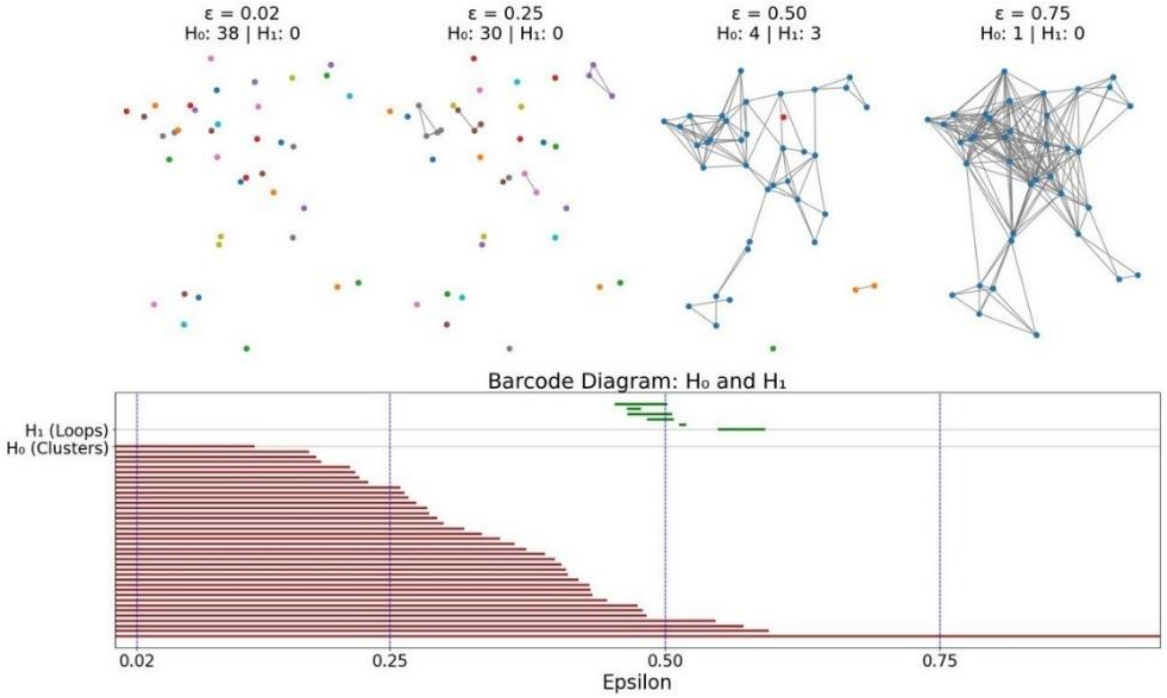

Figure 1. The Topological Evolution and Persistence Barcode of the Dataset.

3.4. The Persistence Diagram [10]

Assume we have the assumptions of (2.3), for $q = 0, \dots, \delta_{\varepsilon_{inf}} = \dim(VR_{\varepsilon_{inf}})$, the q^{th} persistence diagram is defined to be the set of birth-death pairs of the q^{th}-homology classes across all simplicial complexes of the filtration (1):

$$D^q = \{ (b_i, d_i) \in \mathbb{R}^2 \mid i = 1,2, \dots, p_q \text{ and } b_i < d_i \}$$

where b_i and d_i are the birth and death times of an i^{th}- homology class, respectively, for $i = 1,2, \dots, p_q$ and p_q is the total number of the distinct q^{th}-dimensional homology classes that are born during the filtration process. The persistence diagram is defined to be $D = \bigcup_{q=0}^{\delta_{\varepsilon_{inf}}} D^q$.

3.5. The Volve Field

The present study is based on the Volve Field dataset, released publicly by Equinor through the Volve Data Village initiative in 2018 [17]. Volve is an offshore sandstone reservoir located in the North Sea, predominantly composed of Upper Jurassic Hugin and Sleipner formations. These formations are characterized by high porosity shallow marine sandstones with interbedded shales, providing an ideal setting for acoustic and petrophysical log analysis [18]. DT refers to the time it takes for an acoustic compressional wave to travel through a unit length of formation and is typically expressed in microseconds per meter (μs/m). Geologically, DT is influenced by lithology, porosity, fluid saturation, and rock compaction.

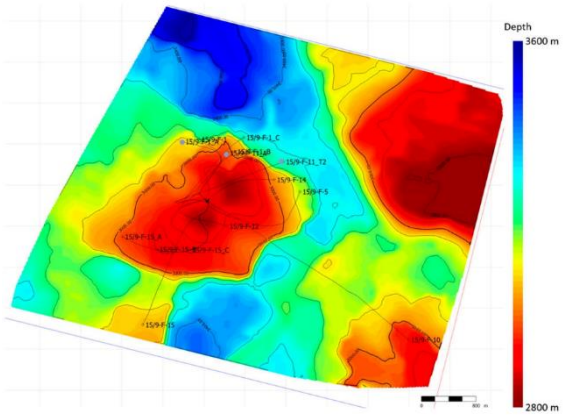

Figure 2. Structural depth map of the reservoir showing well locations and depth contours. Wells used in the core analysis (individual, pairwise, and trio) are in the western flank, within the structurally elevated region. This geographic coherence contributes to topological similarity among them. Depth increases toward the northeast corner.[17]

The Wolf Formation, a heterogeneous siliciclastic reservoir, is characterized by lateral facies variability, diagenetic transformations, and stratigraphic discontinuities. These geological complexities lead to variation in data quality across wells both in terms of depth coverage and signal resolution often reflecting operational goals or drilling constraints. Preliminary inspection Table 1. revealed substantial variation in depth ranges and sampling sizes across the field. For instance, well 15/9-F-11_T2 spans a deep interval from 257 m to 3397.9 m, yielding over 19,400 data points, while 15/9-F-11_A covers approximately 673 m with 11,464 points. In contrast, 15/9-F-1_A, although spatially aligned with the others, spans ~767 m and contributes 10,224 readings. This variation reflects not only geological heterogeneity but also differing acquisition strategies across the reservoir. For in-depth analysis (see Figure 3.), we selected the three most data-rich wells (15/ 9-F-11 T2, 15/ 9-F-11 A, and 15/9-F-1 A) due to their high sampling density and geographic co-location within a unified structural domain. Their spatial proximity ensures minimal lateral facies drift, allowing for robust topological comparisons within a geologically coherent zone.

Table 1. Summary of Well Depth Intervals and Log Sizes in the Volve Field Dataset.

Well Name	Min	Max	Depth Range (m)	Well Size
15/9-F-11 T2	257	3397.919	3140.919	19433
15/9-F-11 A	2453.01	3126.404	673.394	11464
15/9-F-1 A	2472.27	3239.638	767.368	10224
15/9-F-1	223	3330.306	3107.306	9822
15/9-F-15 A	216.8738	3205.477	2988.603	9361
15/9-F-15	145.0848	3171.464	3026.379	7146
15/9-F-1 B	2469.78	3259.79	790.01	3262

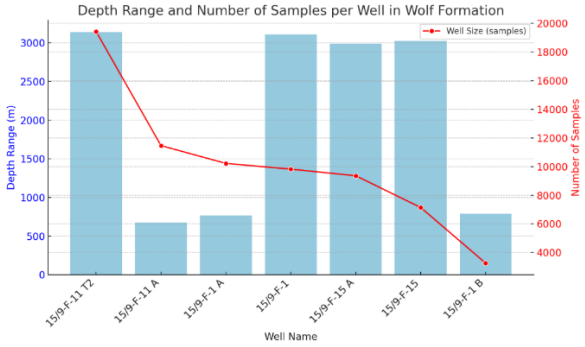

Figure 3: Depth Range and Number of Samples per Well in Wolf Formation

Among these, well (15/ 9-F-11 A) presented distinctive topological behavior when analyzed individually. Although data-rich and structurally aligned, it exhibited mild vertical imbalance in sampling density, particularly within specific depth segments, which initially reduced the clarity of topological descriptors. However, rather than excluding it, we retained (15/ 9-F-11 A) to demonstrate the sensitivity of the framework to localized acquisition variability. Its inclusion ultimately validated the method's robustness: upon pairing with structurally consistent wells, the full topological-predictive alignment was recovered, reaffirming the method's adaptability in heterogeneous reservoir environments.

3.6. Methodology

We adopted the same topological modeling framework established in our previous study [5], which integrates persistent homology with marginal contribution analysis and predictive modeling to assess feature informativeness. The full methodological pipeline comprising log cleaning, anomaly removal via isolation forest, subsampling, landmark selection, persistent homology computation, and topological metric aggregation is detailed in [5] and is briefly summarized in Figure 4.

In this study, the method was systematically applied across eight distinct data scenarios to evaluate stability under varying spatial granularities, including:

1. Full reservoir (7 Wells).
2. Three-Well cluster.
3. Three two-Well combinations.
4. Three Wells.

Missing log values were imputed through depth-wise linear interpolation, while intervals with more than 10% missing coverage were excluded. All logs were standardized to zero mean and unit variance prior to analysis to ensure comparability across features.

Persistent homology was computed using Vietoris–Rips filtrations up to homology dimension 2 with a resolution of $\Delta\varepsilon = 0.01$. A stratified MaxMin landmark selection strategy was applied, typically retaining ~ 30 landmarks per depth segment to preserve geological representativeness while reducing computational costs. From each persistence diagram, four interpretable Betti-0 descriptors (sum, mean, variance, and maximum), were extracted following the definitions formalized in [5] as follows:

let $\ell = (\ell_1, \ell_2, \ldots, \ell_{p_q})$ be the set of all finite lifetimes of Betti-0 features obtained from the persistence diagram, where p_q is the total number of distinct q-dimensional homology classes that are born during the Vietoris–Rips filtration. The q^{th} Betti lifetime metrics for C are defined as follows:

1. The Total sum of q^{th} Betti lifetime: $\beta_q^{sum} = \sum_{i=1}^{p_q} \ell_i$ … (3)
2. The Mean of q^{th} Betti lifetime: $\beta_q^{mean} = \frac{1}{p_q}\sum_{i=1}^{p_q} \ell_i$ … (4)
3. The Variance of q^{th} Betti lifetimes: $\beta_q^{var} = \frac{1}{p_q}\sum_{i=1}^{p_q}(\ell_i - \beta_q^{mean})^2$ … (5)
4. The Maximum of q^{th} Betti lifetime: $\beta_q^{max} = Max_{i=1}^{p_q}\{\ell_i\}$ … (6)

These descriptors quantify the persistence and spread of connected components and serve as model-agnostic indicators of log complexity.

To estimate per-log contribution, we employed linear regression over the inclusion matrix, and in parallel, feature importance scores were computed using a Random Forest model trained on the same subsets. Finally, Pearson

correlation coefficients were calculated between Betti-0 metrics and model-based importance scores, providing a robust measure of alignment across all scales.

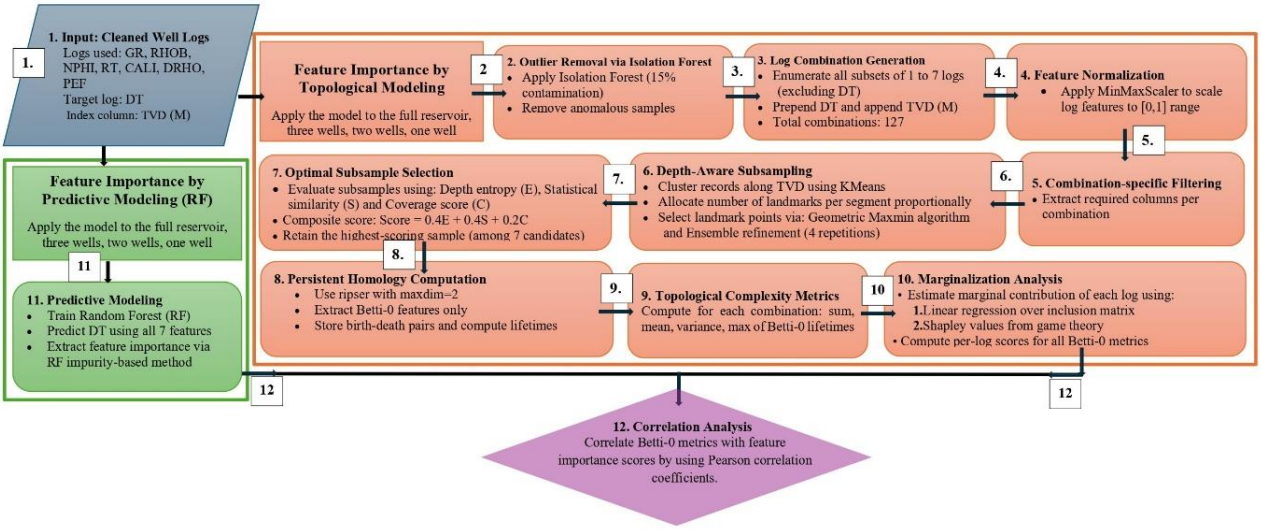

Figure 4. Workflow Diagram: Feature Importance by Using Persistent Homology.

4- Results and Disscusion Section

4.1. Justification for Focusing on β_0 Topological Features

Although both zeroth (β_0) and first (β_1) persistent Betti numbers were computed, the final analysis focused exclusively on β_0 lifetimes. This decision is supported by both visual and statistical evidence. As illustrated in Figure 5, β_1 lifetimes decayed almost instantly across all configurations, while $\beta_0$lifetimes showed broader and more stable persistence ranges. The quantitative summary in Table 2 further reinforces this observation: β_0 lifetimes exhibited a mean of 0.22, a median of 0.20, and maximum values exceeding 0.50, with only ~6% falling below 0.05. In contrast, β_1 lifetimes had a mean of 0.04, a median of 0.03, and maximum values of only 0.09, with nearly 68% shorter than 0.05. Such distributions indicate that β_1 features mostly represent transient cycles or noise, whereas β_0 descriptors capture robust structural components.

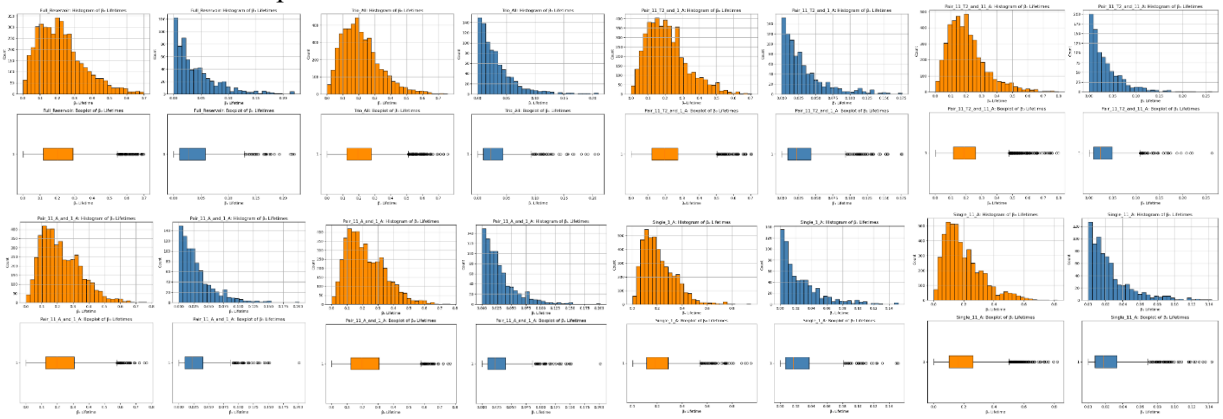

Figure 5. Comparative distribution and spread of Betti-0 and Betti-1 lifetimes across all well configurations.

Table 2. Statistical summary of β_0 and β_1 lifetimes across the full-reservoir configuration

Metric	β_0	β_1
Mean Lifetime	0.22	0.04
Median Lifetime	0.20	0.03
Max Lifetime	0.53	0.09
% Lifetimes < 0.05	5.7%	68%

Moreover, preliminary comparisons confirmed that including β_1 metrics did not alter feature-importance rankings, which remained essentially identical to those derived from β_0 descriptors alone. To maximize

interpretability, clarity, and computational efficiency, we therefore retained β_0 lifetimes as the primary topological indicators of log complexity.

4.2 Topological Complexity across Well Configurations

The β_0-based topological framework was applied across eight distinct scenarios, ranging from full-reservoir combinations to individual well configurations, to evaluate the consistency of topological complexity as a proxy for log informativeness. For each case, we computed four key descriptors of Betti-0 lifetimes (sum, mean, variance, and maximum) for every input log. These descriptors were then used to derive marginal contribution scores via additive linear regression. The resulting topological importance ranks were compared against feature importance ranks extracted from Random Forest models trained on the same log subsets (see Table 3). In all scenarios, logs with lower β_0^{sum} values indicating simpler topological structure tended to be ranked higher in predictive importance. Notably, NPHI consistently exhibited the lowest β_0 complexity and was among the top-ranked predictors. Conversely, DRHO repeatedly showed the highest β_0^{sum} and the lowest importance, reinforcing a strong inverse relationship between topological complexity and feature relevance.

While minor fluctuations occurred among mid-ranked logs (e.g., GR, RT, PEF), the global ordering remained remarkably stable across different spatial scales. This stability affirms the robustness and interpretability of β_0-based descriptors as model-agnostic indicators of feature significance in geophysical well logging.

Table 3. Summary of Betti-0 Topological Complexity and Predictive Importance Ranks Across Eight Well Configurations. Lower β₀ sum values indicate a simpler topological structure. Columns show Betti-0 descriptors and corresponding Random Forest importance ranks.

Well	β_0^{sum} Rank	Log	β_0^{sum}	β_0^{mean}	β_0^{var}	β_0^{max}	RF Importance Rank
All Wells	1 (lowest β_0^{sum})	NPHI	1.459775	0.042935	0.004034	0.140677	1 (highest importance)
	2	RHOB	1.75212	0.051533	0.004894	0.157336	2
	3	GR	1.8009	0.052968	2.79E-05	0.058105	3
	4	RT	1.930251	0.056772	0.009269	0.209125	4
	5	DRHO	2.275403	0.066924	0.003696	0.144487	7 (lowest importance)
	6	CALI	2.606006	0.076647	0.001509	0.134483	6
	7 (highest β_0^{sum})	PEF	2.668875	0.078496	0.002106	0.123131	5
Three Wells	1 (lowest β_0^{sum})	NPHI	1.438811	0.039967	0.003831	0.163556	1 (highest importance)
	2	RHOB	1.498421	0.041623	0.003475	0.157334	3
	3	RT	1.92217	0.053394	0.001218	0.108559	4
	4	GR	2.12327	0.05898	0.002358	0.08175	2
	5	PEF	2.331793	0.064772	0.003799	0.157029	5
	6	CALI	2.620698	0.072797	0.00272	0.153715	6
	7 (highest β_0^{sum})	DRHO	2.979281	0.082758	0.002458	0.125002	7 (lowest importance)
Pair: 11_T2 and 1_A	1 (lowest β_0^{sum})	NPHI	1.580917	0.042727	0.004294	0.14533	1 (highest importance)
	2	RHOB	1.649632	0.044585	0.004327	0.158481	2
	3	GR	1.788914	0.048349	0.004771	0.106913	3
	4	RT	2.158729	0.058344	-0.00103	0.085962	4
	5	PEF	2.31711	0.062625	0.00268	0.119926	5
	6	CALI	2.793037	0.075487	0.002439	0.162376	6
	7 (highest β_0^{sum})	DRHO	2.844403	0.076876	0.00197	0.137446	7 (lowest importance)
Pair: 11_T2 and 11_A	1 (lowest β_0^{sum})	NPHI	1.285188	0.0357	0.002705	0.113821	1 (highest importance)
	2	GR	1.460786	0.040577	0.003224	0.156472	2
	3	RHOB	1.539834	0.042773	0.004157	0.144907	3
	4	RT	1.988042	0.055223	0.00258	0.139212	4

Well	β_0^{sum} Rank	Log	β_0^{sum}	β_0^{mean}	β_0^{var}	β_0^{max}	RF Importance Rank
	5	PEF	2.23208	0.062002	0.003551	0.147703	5
	6	CALI	2.667939	0.074109	0.00219	0.146171	6
	7 (highest β_0^{sum})	DRHO	3.134159	0.08706	0.0039	0.132929	7 (lowest importance)
Pair: 11_A and 1_A	1 (lowest β_0^{sum})	NPHI	1.462917	0.041798	0.003748	0.140983	1 (highest importance)
	2	RHOB	1.496294	0.042751	0.003383	0.152072	2
	3	RT	1.808445	0.05167	0.003242	0.157144	4
	4	GR	2.216519	0.063329	0.00113	0.055476	3
	5	PEF	2.317206	0.066206	0.002547	0.155647	5
	6	CALI	2.72387	0.077825	0.006051	0.153046	6
	7 (highest β_0^{sum})	DRHO	2.905586	0.083017	0.001838	0.120823	7 (lowest importance)
Single: 11_T2	1 (lowest β_0^{sum})	GR	1.284334	0.037775	0.009279	0.220695	2
	2	NPHI	1.297193	0.038153	0.003196	0.1164	1 (highest importance)
	3	RHOB	1.452053	0.042707	0.003152	0.137198	3
	4	RT	1.580487	0.046485	-0.00086	0.088284	4
	5	PEF	2.115838	0.062231	0.004827	0.167658	5
	6	CALI	2.588147	0.076122	0.005295	0.16984	6
	7 (highest β_0^{sum})	DRHO	2.746567	0.080781	0.003048	0.127654	7 (lowest importance)
Single: 1_A	1 (lowest β_0^{sum})	GR	1.197553	0.033265	-0.00148	0.004539	3
	2	NPHI	1.622065	0.045057	0.003691	0.13336	2
	3	RHOB	1.824765	0.050688	0.007582	0.222062	1 (highest importance)
	4	RT	1.979028	0.054973	0.008122	0.351403	4
	5	PEF	2.191724	0.060881	0.004676	0.133681	5
	6	CALI	2.384447	0.066235	0.005739	0.147797	6
	7 (highest β_0^{sum})	DRHO	3.025239	0.084034	0.005176	0.218627	7 (lowest importance)
Single: 11_A	1 (lowest β_0^{sum})	NPHI	1.218699	0.032938	0.003475	0.144373	1 (highest importance)
	2	GR	1.47511	0.039868	0.003773	0.158946	2
	3	RHOB	1.656932	0.044782	0.004993	0.189295	3
	4	RT	1.850922	0.050025	0.002936	0.108929	4
	5	PEF	2.03852	0.055095	0.002824	0.111667	6
	6	CALI	2.639792	0.071346	0.004125	0.131562	5
	7 (highest β_0^{sum})	DRHO	3.038267	0.082115	0.003382	0.147731	7 (lowest importance)

To complement the quantitative ranking comparison, we generated scatter plots (Figure 6) illustrating the alignment between topological simplicity and predictive importance across all eight well configurations. In these plots, the x-axis represents the ascending order of β_0^{sum} values (indicating increasing topological complexity), while the y-axis captures the descending order of Random Forest feature importance. An ideal inverse relationship manifests as a diagonal trend from top-left to bottom-right. Remarkably, this pattern was preserved across all configurations. NPHI and RHOB consistently appeared in the upper-left quadrant, highlighting their simple topological structures and high predictive relevance. DRHO, on the other hand, occupied the bottom-right corner in nearly all cases, reaffirming its complex structure and limited informativeness. Minor deviations were observed in mid-ranked logs such as GR, RT, and PEF, particularly within the single-well scenarios. In these cases, local sampling imbalances may have introduced slight perturbations in the topological signal. Nonetheless, the global trend remained stable, visually reinforcing the validity of Betti-0 descriptors as scalable, model-independent indicators of feature relevance.

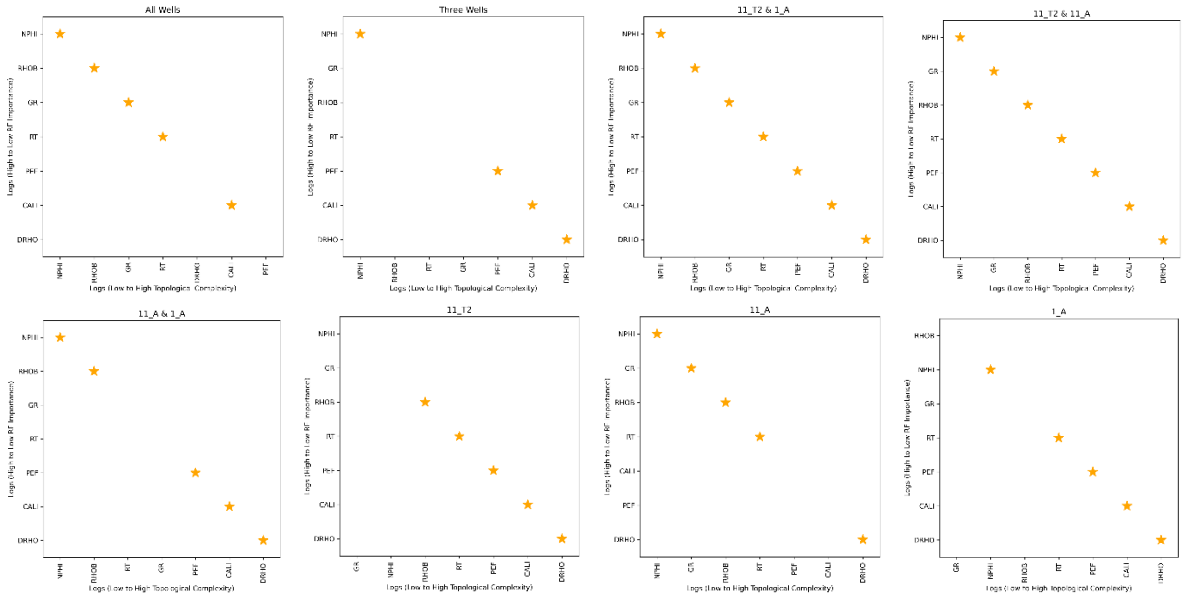

Figure 6. Scatter plots show the inverse relationship between Betti-0 sum rankings (x-axis) and Random Forest feature importance rankings (y-axis) for each of the eight well configurations. Diagonal alignments from top-left to bottom-right indicate strong agreement between topological simplicity and predictive relevance.

4.3 Predictive Importance Scores and Model Performance

To evaluate the consistency of model-based feature relevance across spatial scales, a separate Random Forest regressor was trained for each of the eight well configurations using the full set of seven input logs. Model performance was assessed via the coefficient of determination (R^2), while feature importance scores were extracted using impurity-based criteria. As summarized in Table 4, all models achieved high predictive accuracy, with R^2 values exceeding 0.96 across all configurations. The best performance was observed in the Single: 11_T2 case (R^2 = 0.9845), followed closely by Pair: 15/ 9-F-11 T2 and 15/9-F-1 A (R^2 = 0.9812). Even the more constrained single-well models, such as 15/9-F-1 A, retained strong predictive power (R^2 = 0.9630), confirming the learnability of sonic transit time from the available logs.

Table 4. R^2 Scores of Random Forest Models Across All Well Configurations. All models achieved high predictive performance (R^2 > 0.95), confirming the learnability of DT from input logs.

Case	R^2 Score
Full Reservoir	0.9506
Three Wells	0.9748
Pair: 11_T2 and 1_A	0.9812
Pair: 11_T2 and 11_A	0.9777
Pair: 11_A and 1_A	0.9637
Single: 11_T2	0.9845
Single: 1_A	0.963
Single: 11_A	0.9772

Importantly, the extracted importance scores exhibited strong alignment with the β_0-based topological rankings (see Figure 6.), further validating the proposed method. While minor deviations were present among mid-ranked logs (e.g., GR, RT, PEF), the topological and statistical evaluations remained closely coupled, especially for consistently extreme logs such as NPHI and DRHO.

4.4 Robust Correlation Between Topological Simplicity and Predictive Importance Across Well Configurations

To rigorously assess the relationship between topological simplicity and model-derived feature relevance, we computed Spearman correlation coefficients between the Betti-0 sum rankings and Random Forest feature importance scores across all eight well configurations. These correlations capture the monotonic association between the two measures without assuming linearity. As shown in Table 3, all configurations exhibited strong and

statistically significant inverse correlations ($\rho \leq -0.85,\ p < 0.015$), with perfect negative alignment ($\rho = -1.0$) observed in both Pair (15/9-F-11 T2 and 15/9-F-1 A) and (15/9-F-11 T2 and 15/9-F-11 A). Even in the most constrained single-well scenarios, such as 15/9-F-1 A and 15/9-F-11 T2, the inverse relationship was preserved with high statistical confidence. In the case of well 15/9-F-1 A, the Spearman coefficient reached ($\rho = -0.85714, p = 0.0137$), which, while slightly lower than multi-well configurations, remains highly significant. Upon further inspection, this relative drop was attributed to depth-specific irregularities in landmark distribution. As illustrated in the landmark heatmap (Figure 7.G.). Such imbalances, likely caused by acquisition artifacts or formation inconsistencies, led to reduced representativeness of topological features across depth. Despite these challenges, the topological descriptors still aligned strongly with predictive importance, underscoring the method's robustness even under suboptimal sampling. We further validated these findings by computing Pearson correlation coefficients between the Betti-0 sum and Random Forest importance rankings. The results confirmed that topological simplicity reliably predicts log relevance across spatial scales, independent of data size or well configuration. Combined with the high predictive R^2 values (Table 5), these correlations reinforce the generalizability and interpretability of the β_0-based topological framework.

Table 5. Spearman Correlation of β₀ Sum Rankings with Random Forest Feature Importance. Strong negative correlations across most cases indicate that lower topological complexity is consistently associated with higher predictive relevance.

Case	Spearman Correlation	*p*-value
All Wells	-0.85714	0.01370
Three Wells	-0.89286	0.00681
15/9-F-11T2 and 15/9-F-1 A	-1	0
15/9-F-11 T2 and 15/9-F-11 A	-1	0
15/9-F-11 A and 15/9-F-1 A	-0.96429	0.00045
15/9-F-11 T2	-0.96429	0.00045
15/9-F-1 A	-0.85714	0.01370
15/9-F-11 A	-0.96429	0.00045

4.5 Behavior of Well 15/9-F-1 A and Depth-Wise Landmark Effects:

Among the analyzed configurations, well 15/9-F-1 A exhibited the lowest Spearman correlation among all scenarios ($\rho = -0.85714, p = 0.0137$). While still statistically significant, this relative drop prompted a closer examination of the well's internal topological structure and sampling distribution. Geologically, well 15/9-F-1 A spans an intermediate depth range of approximately 767 meters (from 2472.27 to 3239.64 m) and contains 10,224 log readings. Despite this dense sampling profile, the well lies within a transitional depositional environment and such settings are prone to facies mixing, lateral heterogeneity, and diagenetic irregularities, which can reduce the structural clarity of recorded logs. The depth-wise landmark distribution heatmap (Figure 7.I.) revealed a notable vertical imbalance: segment 7 was sparsely sampled with as few as one or two landmarks, whereas segments 5 and 6 exhibited oversaturation. This uneven distribution likely stemmed from acquisition limitations or local formation properties, leading to biased persistence descriptors and reduced topological resolution in underrepresented zones. Complementary β_0 lifetime plots (7.B-C.) confirmed increased variance and component dispersion, while the 3D scatter view (7.A.) placed most configurations of 15/9-F-1 A in a low-correlation, high-variability regime. Despite acceptable coverage scores (7.H.), the alignment between topological descriptors and feature relevance was partially compromised. Remarkably, when well 15/9-F-1 A was paired with 15/9-F-11 T2, a deeper, more geologically stable well with over 19,000 readings, the correlation with feature importance improved to a perfect inverse alignment ($\rho = -1.0$). This suggests that the coherent structural signal from 15/9-F-11 T2 effectively compensated for the internal variability of 15/9-F-1 A.

This case highlights the local sensitivity of topological descriptors to vertical data quality and sampling balance. However, it also affirms the global robustness of the Betti-based framework: even under irregular depth-wise coverage, the method maintained statistical significance and regained full coherence when integrated with structurally consistent wells

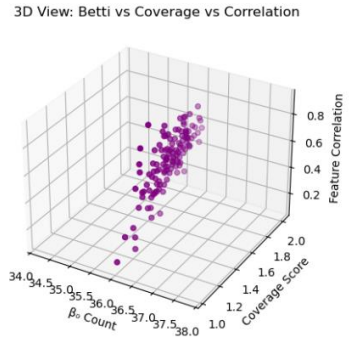

Figure 7.A. 3D View of β_0 Count, Coverage, and Feature.

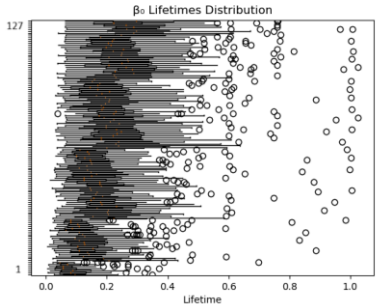

Figure 7.B. β_0 Lifetimes Distribution per Combination (Boxplot).

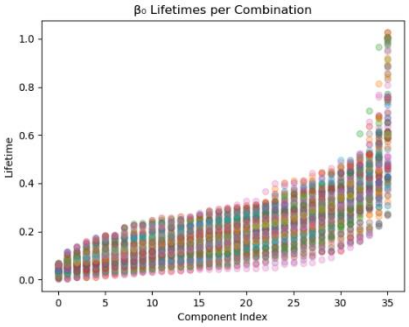

Figure 7.C. β_0 Lifetimes per Combination (Scatter).

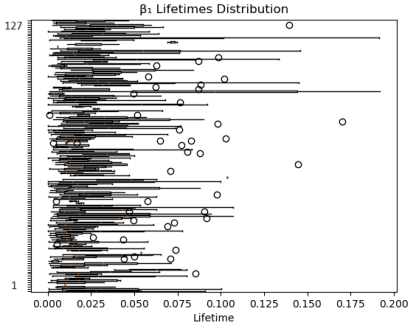

Figure 7.D. β_1 Lifetimes Distribution per Combination (Boxplot).

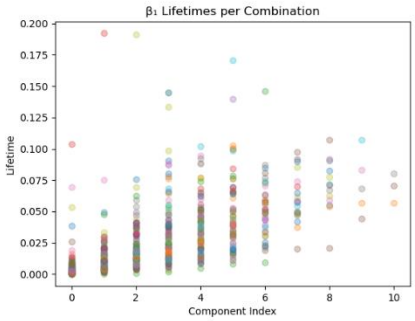

Figure 7.E. β_1 Lifetimes per Combination (Scatter).

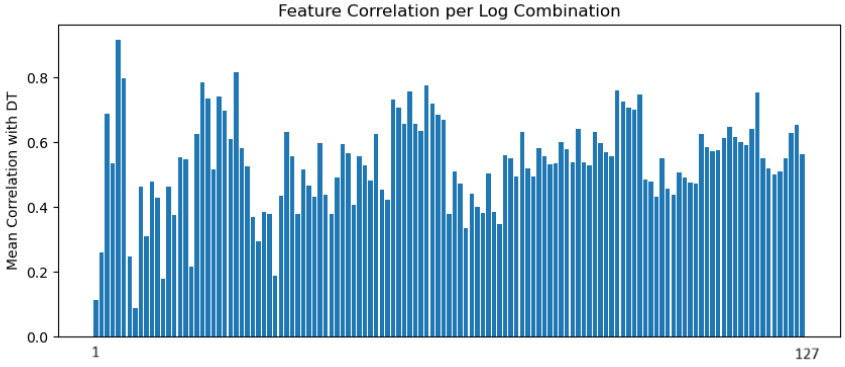

Figure 7.G. Variability of Mean Correlation with DT per Log Combination

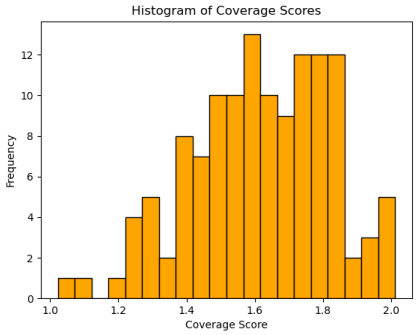

Figure 7.H. Landmark Distribution Across Depth Segments for all Combinations.

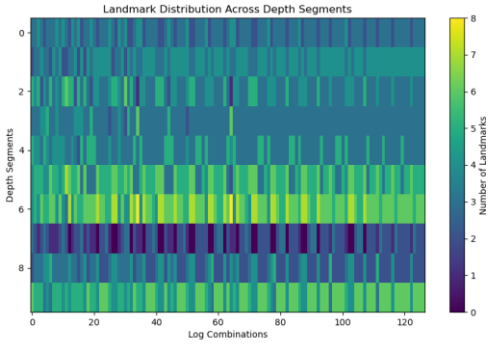

Figure 7.I. Landmark Distribution Across Depth Segments for all Combinations.

Figure 7. Diagnostic Visualizations for Well 15/9-F-1 A

While the framework was demonstrated using the Volve dataset, the methodology is not reservoir specific. By adjusting landmark selection strategies and filtration parameters to match local geological heterogeneity, the approach can be transferred to other fields and logging environments. This enhances its applicability as a general model-agnostic tool for well-log prioritization and feature relevance analysis.

5- Conclusion Section

This study builds upon our prior work by applying the Betti-0-based topological framework to multiple spatially localized well configurations within a reservoir. The objective was to assess whether the inverse relationship between topological simplicity and feature importance previously established at the full-reservoir scale remains stable under more constrained, site-specific data conditions. The results confirm that Betti-0 lifetime metrics are robust indicators of log informativeness, even when data availability is limited to individual wells or partial combinations. Across all eight scenarios, logs with lower β_0 sum consistently ranked higher in Random Forest importance, with Spearman correlations exceeding –0.85 in every case. In the single-well case of 15/9-F-1 A, the slightly lower correlation was attributed to vertical sampling imbalance and localized acquisition gaps. Nonetheless, statistical significance was preserved, and perfect inverse alignment was achieved when 15/9-F-1 A was paired with a structurally consistent neighbor.

Moreover, all predictive models maintained high accuracy ($R^2 > 0.95$), highlighting the learnability of DT from topologically filtered inputs. The consistency of outcomes across full, trio, pairwise, and single-well cases demonstrates the generalizability of topological complexity as a model-agnostic relevance metric. Unlike conventional feature attribution techniques that rely on model internals or performance degradation, our approach directly captures the geometric essence of the input data. This makes it especially valuable in scenarios involving limited supervision, evolving model architecture, or high-dimensional logs.

Although validated on the Volve dataset, the framework is not reservoir-specific. With appropriate adjustment of landmark sampling and filtration parameters, it can be extended to other reservoirs and logging environments.

Ultimately, this localized extension of topological analysis offers a principled, interpretable, and scalable solution for log selection and predictive screening in geophysical machine learning. Future applications may integrate this framework into real-time logging workflows and adaptive acquisition strategies, where rapid, data-efficient, and model-independent assessments of feature utility are essential.

Conflicts Of Interest

The authors declare no conflicts of interest.

Funding

This work was supported by the College of Science, University of Basrah, as part of the postgraduate research program. No specific grant was assigned.

Acknowledgment

The authors would like to thank Semaa Alessa from the University of Houston for her help and understanding in the field of petrophysics. Her ideas were helpful in figuring out what the data meant and making sure that the analysis was useful for describing subsurface reservoirs.

References

[1] **A. McDonald**, "Data Quality Considerations for Petrophysical Machine-Learning Models," *Petrophysics*, vol. 62, no. 6, pp. 585–613, Dec. 2021, doi: 10.30632/PJV62N6-2021a1.

[2] **H. Wang, Y. Wu, Y. Zhang, F. Lai, Z. Feng, B. Xie, et al.**, "Uncertainty and explainable analysis of machine learning model for reconstruction of sonic slowness logs," *Artificial Intelligence in Geosciences*, vol. 4, pp. 182–198, Dec. 2023, doi: 10.1016/J.AIIG.2023.11.002.

[3] **C. Nero, A. A. Aning, S. K. Danuor, and V. Mensah**, "Prediction of compressional sonic log in the western (Tano) sedimentary basin of Ghana, West Africa using supervised machine learning algorithms," *Heliyon*, vol. 9, no. 9, p. e20242, Sep. 2023, doi: 10.1016/J.HELIYON.2023.E20242.

[4] **A. Fisher, C. Rudin, and F. Dominici**, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *J Mach Learn Res*, vol. 20, p. 177, Dec. 2019. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8323609/

[5] **S. K. Hassan and H. M. Ali**, "Persistent Homology for Geophysical Insights: A Topological Metric for Feature Importance," *Basrah Journal of Science*, "preparint", 2025.

[6] **L. Breiman**, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[7] **A. Altmann, L. Tolo¸si, T. Tolo¸si, O. Sander, and T. Lengauer**, "Data and text mining Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[8] **C. Molnar**, "Interpretable Machine Learning A Guide for Making Black Box Models Explainable," 2020, Available: https://christophm.github.io/

[9] **S. M. Lundberg, P. G. Allen, and S.-I. Lee**, "A Unified Approach to Interpreting Model Predictions," *Adv Neural Inf Process Syst*, vol. 30, 2017, Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[10] **Herbert. Edelsbrunner and J. . Harer**, "Computational topology : an introduction." *American Mathematical Society*, 2010. Available: https://search.worldcat.org/title/700185732.

[11] **G. Carlsson**, "Topology and Data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009, doi:10.1090/S0273-0979-09-01249-X.

[12] **M. Hajij, G. Zamzmi, and F. Batayneh**, "TDA-Net: Fusion of Persistent Homology and Deep Learning Features for COVID-19 Detection from Chest X-Ray Images," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2021, pp. 4115–4119, Jan. 2021, doi: 10.1109/EMBC46164.2021.9629828.

[13] **L. M. Alhelfi and H. M. Ali**, "Using Persistence Barcode to Show the Impact of Data Complexity on the Neural Network Architecture," *Iraqi Journal of Science*, vol. 63, no. 5, pp. 2262–2278, May 2022, doi: 10.24996/IJS.2022.63.5.37.

[14] **B. J. Stolz**, "Outlier-Robust Subsampling Techniques for Persistent Homology," *Journal of Machine Learning Research*, vol. 24, pp. 1–35, 2023, Accessed: May 10, 2025, Available: http://jmlr.org/papers/v24/21-1526.html.

[15] F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman, "Subsampling Methods for Persistent Homology," *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, pp. 2133–2141, Jun. 2014, Available: https://arxiv.org/abs/1406.1901v1.

[16] **W. H. Guss and R. Salakhutdinov**, "On Characterizing the Capacity of Neural Networks using Algebraic Topology," Feb. 2018, doi: https://doi.org/10.48550/arXiv.1802.04443.

[17] **Equinor**, "Volve field data set download - Equinor." Accessed: Jun. 20, 2025, Available: https://www.equinor.com/energy/volve-data-sharing.

[18] **D. Pelemo-Daniels and R. R. Stewart**, "Petrophysical Property Prediction from Seismic Inversion Attributes Using Rock Physics and Machine Learning: Volve Field, North Sea," *Applied Sciences 2024,* vol. 14, no. 4, p. 1345, Feb. 2024, doi: 10.3390/APP14041345.