

Coreless Geo: A Machine Learning and Retrieval-Augmented Generation Framework for Geomechanical Rock Property Prediction and Literature-Grounded Geological Interpretation

Mrunaal Gathe, Parth Karegaonkar

¹²Department of Artificial Intelligence and Data Science, Zeal College of Engineering and Research, Narhe, Pune – 411041, Maharashtra, India

³Project Guide, Department of Artificial Intelligence and Data Science, Zeal College of Engineering and Research, Narhe, Pune – 411041, Maharashtra, India

Email: mruaalg@gmail.com, parthkaregaonkar@gmail.com

Abstract — *Coreless Geo is an end-to-end artificial intelligence system that predicts four critical geomechanical rock properties — Young’s Modulus (E), Poisson’s Ratio (ν), Unconfined Compressive Strength (UCS), and Brittleness Index (BI) — directly from well log measurements, eliminating dependence on expensive physical core sampling. The framework benchmarks three regression algorithms, Random Forest, XGBoost, and Support Vector Regression, and selects the best-performing model per property using Leave-One-Well-Out (LOWO) cross-validation, an evaluation protocol that yields an honest estimate of cross-well generalization rather than the optimistically biased scores produced by random train–test splits. SHAP (SHapley Additive exPlanations) is applied to expose which well log features drive each prediction, addressing the black-box criticism levelled at prior ML-based geomechanical studies. A Retrieval-Augmented Generation (RAG) layer, built on a ChromaDB vector store of curated geology literature and the Claude API, converts raw numerical predictions into structured, citation-grounded geological interpretation reports covering lithology, estimated formation age, stress regime, and drilling recommendations. The complete pipeline is deployed as an interactive Streamlit web application that allows non-technical petroleum engineers to obtain predictions, explanations, and interpretive reports within seconds of uploading a well log CSV file. A dedicated cross-basin generalization experiment, applying a North Sea-trained model to a geologically distinct basin without retraining, explicitly quantifies the domain-shift penalty that is rarely examined in the existing literature. Coreless Geo thereby contributes a reproducible, explainable, and deployable methodology that bridges data-driven property prediction with domain-grounded geological reasoning.*

Keywords — *Machine Learning; Well Log Analysis; Geomechanical Property Prediction; Retrieval-Augmented Generation; SHAP Explainability; XGBoost; Cross-Basin Generalization; Leave-One-Well-Out Cross-Validation; Geological Interpretation.*

1. INTRODUCTION

1.1 Background

Geomechanical rock properties — Young’s Modulus (E), Poisson’s Ratio (ν), Unconfined Compressive Strength (UCS), and Brittleness Index (BI) — are critical inputs for petroleum engineering decisions, including hydraulic fracturing design, wellbore stability analysis, and reservoir management. Traditionally, these properties are determined through laboratory analysis of physical core samples extracted from the wellbore during drilling. Core acquisition, however, costs between USD 500 and USD 1,500 per metre, is time-consuming, and is not performed for every well in a field; a typical oilfield operator may drill 200 wells but core only 10–15 of them, leaving the majority of wells with incomplete geomechanical characterization.

Coreless Geo addresses this industry-wide problem by training supervised machine learning models on wells where both well log data and core measurements exist, and then deploying those models to predict geomechanical properties for wells

that were never cored. Well logs — continuous measurements of gamma ray, bulk density, neutron porosity, resistivity, and sonic transit time — are available for virtually every drilled well and are inexpensive to acquire. Once the log-to-property relationship is learned from cored wells, it can, in principle, be applied at any depth in any uncored well within the same geological basin. The present work (v2.0) extends an earlier exploratory framework with four substantive additions: rigorous Leave-One-Well-Out (LOWO) cross-validation for honest generalization assessment, SHAP explainability for model transparency, a Retrieval-Augmented Generation (RAG) interpretation layer grounded in published geological literature, and a production-ready Streamlit web application. Together, these additions transform the system from an experimental notebook into a deployable, interpretable, and academically defensible geomechanical prediction pipeline.

1.2 Problem Statement

Core acquisition in petroleum engineering is expensive, spatially sparse, and frequently unavailable for historical or low-priority wells. Existing machine learning approaches to geomechanical property prediction from well logs suffer from three unresolved limitations:

- Validation bias: Models are typically validated using random 80/20 splits drawn from the same wells used for training, which produces optimistically biased R^2 scores that do not reflect real-world generalization to genuinely new wells.
- Lack of interpretive context: Predicted values (e.g., UCS = 42.1 MPa) are presented as bare numbers without geological interpretation, limiting their usefulness to domain experts and making results inaccessible to non-technical stakeholders.
- Untested cross-basin applicability: Whether a model trained on data from one basin (e.g., the North Sea) can be meaningfully applied to wells in a geologically distinct basin (e.g., an Indian sedimentary basin) is rarely examined, yet this question is central to practical deployment in data-scarce regions.

Coreless Geo directly addresses all three limitations through LOWO cross-validation, RAG-grounded geological report generation, and an explicit cross-basin generalization analysis.

1.3 Motivation

The petroleum industry generates vast quantities of well log data for every drilled well, yet geomechanical properties remain unknown for the majority of wells owing to the cost of coring. This creates a persistent gap between available data and actionable knowledge. At the same time, the rapid advancement of machine learning and large language models (LLMs) has created an unprecedented opportunity to close this gap: ML models can learn log-to-property mappings from the limited set of cored wells, while LLMs grounded in geological literature can transform numerical predictions into the structured, contextualized reports that geologists and engineers actually use for decision-making.

India presents a particularly compelling motivation for this line of work. The Krishna–Godavari Basin, Cambay Basin, and Mumbai Offshore region are active petroleum provinces in which geomechanical data is sparse relative to the volume of available well log data. A system that can reliably predict and interpret geomechanical properties from logs — with honest uncertainty characterization and literature-grounded interpretation — could directly support Indian petroleum operations at a fraction of the cost of an expanded core acquisition programme.

1.4 Research Objectives

The principal aim of this work is to design and evaluate an end-to-end system that predicts geomechanical rock properties from well log data under an honest validation protocol, provides feature-level explainability, generates literature-grounded geological interpretation via RAG, and delivers all of this functionality through a deployable web application. The specific objectives are to:

- Train and benchmark Random Forest, XGBoost, and Support Vector Regression models on North Sea well log data for all four target properties (E, ν , UCS, BI).

- Implement Leave-One-Well-Out cross-validation to obtain a true measure of generalization performance across wells.
- Integrate SHAP explainability to identify which well log features drive each prediction at each depth.
- Build a RAG pipeline using ChromaDB and the Claude API to generate structured JSON geological interpretation reports grounded in geology literature.
- Conduct a cross-basin generalization experiment comparing North Sea model performance against an independent basin dataset.
- Deploy the complete system as a Streamlit web application supporting CSV upload, prediction, visualization, and report export.

1.5 Novel Contributions

This work makes the following contributions relative to prior literature in ML-based geomechanical property prediction:

- Cross-well generalization validation: LOWO testing provides an honest estimate of model performance on genuinely unseen wells, as opposed to held-out samples drawn from the same wells used in training.
- Physics-aware feature engineering: derived features — acoustic impedance, shale volume, and the velocity ratio V_p/V_s — are computed from raw logs prior to model training, encoding domain knowledge directly into the input feature space.
- RAG-grounded interpretation: predictions are interpreted by an LLM that retrieves relevant passages from a curated database of geology papers and formation reports, producing interpretations accompanied by literature citations.
- Cross-basin domain-shift quantification: the degradation in performance when a model is applied to a geologically distinct basin is explicitly measured and framed as a research finding rather than treated as an incidental limitation.
- Production-ready deployment: a Streamlit application allows non-technical users to obtain predictions, explanations, and reports without writing or running any code.

2. LITERATURE SURVEY

2.1 Overview

The prediction of geomechanical rock properties from well log data has been an active research area since the early 2000s. Early approaches relied on empirical correlations — deterministic equations relating sonic travel time or bulk density to rock strength — developed for specific geological formations. Correlations such as the Castagna V_p – V_s relationship [11] and the empirical UCS formulas of Chang et al. provide approximate estimates but fail to capture the complex, non-linear, basin-specific relationships between multiple log inputs and mechanical properties.

The machine learning era introduced data-driven models capable of learning these non-linear relationships directly from data. Research has progressed from single-model approaches (SVM, ANN) to ensemble methods such as Random Forest [2] and XGBoost [3], and subsequently to hybrid stacking models. More recently, physics-informed machine learning, in which domain-knowledge constraints are embedded into model training, has emerged as a frontier approach [9]. The use of LLMs and Retrieval-Augmented Generation for geological interpretation remains largely unexplored in the peer-reviewed literature, representing the principal novel contribution of the present work.

2.2 Comparative Analysis of Prior Work

Table 1 summarizes representative prior studies in ML-based geomechanical and petrophysical property prediction, together with the specific limitation that each is found to have, and which Coreless Geo addresses.

Author(s)	Year	Method	Key Contribution	Limitation Addressed by This Work
Ameen et al. [6]	2019	Random Forest, SVM	Predicted UCS and E from well logs with $R^2 > 0.87$	No cross-well validation; no interpretation layer
Al-Anazi & Gates	2010	Support Vector Regression	SVR outperformed linear regression for UCS prediction	Basin-specific; no generalization test
Sattari et al.	2020	XGBoost, Gradient Boosting	Ensemble models improved robustness for multi-property prediction	No explainability; no deployment
Ehsan et al. [7]	2024/25	Gradient Boosting + CWT	ML-optimized pore pressure prediction on a limited dataset, $R^2 = 0.91$	Single property; no RAG interpretation
Zhu et al. [10]	2025	Deep-learning data augmentation	Improved lithology classification via augmented well logs	Classification only; not geomechanics
Ma et al. [9]	2024	Review: hybrid + physics models	Surveyed physics-constrained neural networks for geomechanical parameters	No open-source implementation; no RAG
Wang et al. [8]	2025	ML + SHAP, North Sea F3 block	SHAP used to guide feature selection for well log prediction	Single basin; no cross-basin test
Proposed: Coreless Geo	2025–26	RF / XGBoost / SVR + RAG + SHAP + LOWO	End-to-end prediction, explanation, and literature-grounded interpretation with cross-basin analysis	—

Table 1. Comparative Analysis of Prior Work in ML-Based Geomechanical Property Prediction

2.3 Research Gap Analysis

Table 2 consolidates the gaps identified in the existing literature and maps each gap to the corresponding design element of Coreless Geo.

Identified Gap	How Coreless Geo Addresses It
Models validated on random splits from the same wells, producing overly optimistic R^2	Leave-One-Well-Out (LOWO) cross-validation provides honest cross-well generalization metrics
Predicted numbers carry no geological context or interpretation	RAG pipeline generates structured JSON geological reports with literature citations
Cross-basin applicability rarely tested	Explicit domain-shift experiment: North Sea model evaluated on an independent basin dataset
Black-box predictions with no feature-level explanation	SHAP explainability is computed per prediction and per target property

Identified Gap	How Coreless Geo Addresses It
No deployment beyond exploratory notebooks	Full Streamlit web application with CSV upload, prediction, visualization, and report export
LLM interpretation ungrounded in domain literature	ChromaDB vector database of geology papers grounds every interpretive claim in a retrievable source

Table 2. Research Gap Analysis

3. SYSTEM ARCHITECTURE AND DESIGN

3.1 Design Requirements

The system is required to ingest well log CSV files, automatically preprocess and engineer features, train and benchmark the three candidate regression models, perform LOWO validation, generate depth-wise predictions with SHAP explanations, query the RAG pipeline for geological interpretation, and expose all of this functionality through a browser-based interface without requiring the end user to write code. Table 3 summarizes the key non-functional requirements that constrain the design.

Requirement	Target / Constraint
Prediction latency	A 1,000-row well log CSV must be processed in under 5 seconds on standard CPU hardware
Accuracy	LOWO R^2 must exceed 0.75 for the best model on at least three of four target properties
Reproducibility	Identical input must always yield identical ML predictions (fixed random seeds)
Portability	The application must run on Windows, macOS, and Linux without requiring a GPU
Usability	A petroleum engineer with no Python knowledge must be able to obtain predictions and reports through the web UI alone
Interpretability	Every numerical prediction must be accompanied by SHAP values and a geological interpretation report

Table 3. Non-Functional Requirements

3.2 Overall System Architecture

Coreless Geo uses a layered architecture comprising three distinct processing layers communicating through well-defined interfaces. The architecture is deliberately implemented as a single-process Streamlit application to maximize deployability and minimize infrastructure complexity, while maintaining a clean internal separation of concerns. The three layers are:

- ML Prediction Layer: scikit-learn and XGBoost models trained on well log data, wrapped in a preprocessing pipeline. Responsible for all numerical predictions and runs entirely on CPU.
- RAG Interpretation Layer: a ChromaDB vector database of geology documents, sentence-transformer embeddings, and the Claude API. Responsible for converting numerical predictions into grounded geological reports.
- Presentation Layer: a Streamlit web application that orchestrates both layers, handles file upload and download, renders visualizations, and exposes all functionality through a browser interface.

3.3 ML Pipeline Architecture

The ML pipeline follows a strict sequential flow: (i) raw CSV ingestion via pandas; (ii) column standardization, mapping user column names to internal standard names (GR, RHOB, NPHI, DT, RT, DEPTH); (iii) missing-value imputation using a median-per-feature strategy; (iv) derived-feature computation — acoustic impedance ($AI = RHOB \times Vp$), shale volume (VSH), and Vp/Vs ratio from DT; (v) feature scaling using StandardScaler fitted on training data only and applied to test data; (vi) independent training of Random Forest, XGBoost, and SVR, each with randomized hyperparameter search (RandomizedSearchCV, 50 iterations, 5-fold CV); (vii) LOWO validation, where an outer loop iterates over wells, trains on $N-1$ wells, tests on the held-out well, and records R^2 and RMSE per property per well; (viii) model selection, where the best model per property is selected by median LOWO R^2 ; and (ix) SHAP computation using TreeExplainer for Random Forest and XGBoost, and KernelExplainer for SVR. The pipeline outputs a depth-indexed DataFrame containing the four predicted properties together with their SHAP values.

3.4 RAG Pipeline Architecture

The RAG pipeline operates in two phases: an offline build phase executed once, and an online query phase executed at prediction time.

Offline build phase: all PDF and text documents in the `/geology_docs/` directory are loaded using LangChain's PyPDFLoader; documents are split into 500-word chunks with 50-word overlap using a RecursiveCharacterTextSplitter; each chunk is embedded using sentence-transformers/all-MiniLM-L6-v2 (a free, local embedding model with no API cost); and the resulting embeddings are stored in a persistent ChromaDB vector store.

Online query phase (executed per prediction): a natural-language search query is constructed from the predicted property values and log statistics; the top-3 most relevant chunks are retrieved from ChromaDB via cosine similarity search; a structured prompt combining well log values, ML predictions, and retrieved context is assembled; the prompt is sent to the Claude API (claude-sonnet-4-6 model) with a request for JSON-formatted output; the JSON response is parsed and validated; and the structured report is returned as a Python dictionary for display and export.

3.5 End-to-End Application Workflow

The complete user workflow proceeds as follows: the user opens the application in a browser; navigates to the Training tab and uploads a multi-well CSV containing log features and target properties; the system preprocesses the data, trains the three candidate models, runs LOWO validation, and displays a benchmark table; the user then navigates to the Prediction tab and uploads an unseen well log CSV (no target columns required); the system applies the best model per property, generates depth-profile predictions, and renders log plots; SHAP values are computed and displayed as feature-importance bar charts and per-depth waterfall plots; the RAG pipeline is queried and a structured geological report is generated and displayed as interpretive cards; and finally the user can download the predictions as CSV, the SHAP plots as PNG images, and the geological report as a PDF.

4. DATASET, PREPROCESSING, AND FEATURE ENGINEERING

4.1 Dataset Description and Source

Coreless Geo uses well log data from the Volve oilfield dataset (North Sea, Norway), which was publicly released by Equinor in 2018 [13]. This is one of the most comprehensive open petroleum datasets available, containing well log measurements and core-derived geomechanical properties from multiple wells within the same geological basin. The dataset is used under the Equinor Open Data licence for academic and research purposes. The Volve dataset is well-suited to this work for three reasons: it contains real core measurements, rather than synthetic data, that serve as ground truth for model training; it includes multiple wells from the same basin, enabling meaningful LOWO cross-validation; and it is geologically representative of well-documented North Sea formations, allowing the RAG layer to retrieve grounded interpretive context.

For the cross-basin generalization experiment, publicly available well log data from the Kansas Geological Survey open database — representing a geologically distinct basin with different lithology and tectonic history — is used to simulate the domain-shift scenario. This enables a quantitative measurement of how much model performance degrades when applied outside the training basin.

4.2 Input and Target Variables

Variable	Type	Unit	Description
DEPTH	Input (index)	m	Measured depth along wellbore
GR	Input feature	API	Gamma Ray — lithology indicator (high = shale)
RHOB	Input feature	g/cc	Bulk density — porosity and strength indicator
NPHI	Input feature	fraction	Neutron porosity — fluid and lithology sensitivity
DT	Input feature	μs/ft	Sonic transit time — inversely related to rock stiffness
RT	Input feature	ohm·m	Deep resistivity — fluid saturation indicator
AI	Derived feature	g/cc × m/s	Acoustic impedance = RHOB × Vp — key strength predictor
VSH	Derived feature	fraction	Shale volume from Gamma Ray normalization
Vp/Vs	Derived feature	dimensionless	Velocity ratio — directly related to Poisson's Ratio
E	Target	GPa	Young's Modulus — rock stiffness
ν	Target	dimensionless	Poisson's Ratio — lateral deformation response
UCS	Target	MPa	Unconfined Compressive Strength — rock strength
BI	Target	dimensionless (0–1)	Brittleness Index — fracturing behaviour

Table 4. Input and Target Variables

4.3 Preprocessing Pipeline

The following preprocessing steps are applied, in fixed sequence, to all data prior to model training or prediction:

- Column name standardization: user column names are mapped to internal standard names via a configurable dictionary, allowing the system to accept datasets from different sources.
- Depth-indexed sorting: data is sorted by DEPTH in ascending order to preserve the stratigraphic sequence.
- Missing-value identification: missing values are flagged using pandas isna(); if a target column has more than 30% missing values, a warning is raised and the corresponding well is excluded from training.
- Feature imputation: missing input-feature values are filled using the column median computed from the training set; the same median values are subsequently applied to test or prediction data.
- Outlier handling: values beyond the 1st and 99th percentile per column are clipped to those bounds to prevent extreme values from distorting model training.
- Feature scaling: all input features are scaled using StandardScaler (zero mean, unit variance); the scaler is fitted exclusively on training data and saved alongside the model for consistent application to prediction data.

4.4 Feature Engineering

Three domain-knowledge-driven derived features are computed from raw log readings prior to model training:

- Acoustic Impedance (AI): $AI = RHOB \times Vp$, where $Vp = 304800 / DT$ (converting from $\mu s/ft$ to m/s). AI is one of the strongest predictors of rock mechanical properties in published petrophysics literature, as it directly encodes both density and velocity, the two physical drivers of stiffness.
- Shale Volume (VSH): computed using the linear Gamma Ray method, $VSH = (GR - GR_{min}) / (GR_{max} - GR_{min})$, clipped to $[0, 1]$. VSH provides a lithology indicator that correlates strongly with UCS and BI, since clay-rich rocks behave fundamentally differently from clean sandstones.
- Velocity Ratio (Vp/Vs): computed as DT_S / DT (where DT_S is shear sonic transit time, if available, otherwise estimated via the Castagna equation [11]). This ratio is directly related to Poisson's Ratio via the exact elastic relationship $v = (Vp/Vs^2 - 2) / (2 \times Vp/Vs^2 - 2)$, providing a physics-informed feature for Poisson's Ratio prediction.

4.5 Data Augmentation Strategy

Given the limited number of wells in any single dataset, data augmentation is applied to expand the effective training set. Gaussian noise injection creates, for each training depth sample, three augmented copies by adding Gaussian noise ($\sigma = 2\%$ of column standard deviation) to all input features while leaving target values unchanged; this triples the effective training set size while remaining within physically plausible ranges. A sliding depth window further augments the training data by creating overlapping 5-point depth windows as additional training rows, with features averaged over the window to introduce local depth-context sensitivity. Augmentation is applied only to training data, never to validation or test data, to prevent data leakage.

4.6 Exploratory Data Analysis

Prior to model training, exploratory data analysis is performed to characterize feature distributions, identify outliers, assess feature–target relationships, and validate that derived features carry the expected physical signal. The analysis comprises: a statistical summary table (mean, standard deviation, minima, maxima, and quartiles per well) used to quantify inter-well variability; a Pearson correlation heatmap across all features and targets, with expected strong correlations between DT and E (negative), RHOB and UCS (positive), AI and E (positive), and VSH and BI (negative); multi-track depth log plots for visual quality control; crossplots of key feature–target pairs colour-coded by well identifier to assess inter-well consistency; distribution analysis (histograms and kernel density estimates) to detect skewness that may warrant log-transformation; and box plots of each target property grouped by well, which directly inform the expected difficulty of LOWO validation. Inter-well variability observed in these box plots indicates that LOWO cross-validation constitutes a meaningfully challenging generalization test, and DT and RHOB are expected to be the strongest individual predictors of E and UCS, consistent with established rock-physics theory.

5. METHODOLOGY: MODEL IMPLEMENTATION AND RAG PIPELINE

5.1 Regression Models Benchmarked

Three supervised regression models are benchmarked independently for each of the four target properties, chosen to cover distinct algorithmic families:

- Random Forest (RF): an ensemble of decision trees that reduces variance through averaging, robust to outliers and noisy features, and naturally provides feature importances. Hyperparameters tuned: `n_estimators`, `max_depth`, `min_samples_split`, `max_features`.
- XGBoost (XGB): gradient-boosted decision trees with L1 and L2 regularization, representing the state of the art for structured/tabular data and offering faster training than RF on large datasets. Hyperparameters tuned: `n_estimators`, `learning_rate`, `max_depth`, `subsample`, `colsample_bytree`, `reg_alpha`, `reg_lambda`.

- Support Vector Regression (SVR): a kernel-based method that fits a hyperplane in a high-dimensional feature space, effective when the number of features is large relative to the number of samples. Hyperparameters tuned: C, epsilon, kernel (RBF), gamma.

5.2 Leave-One-Well-Out (LOWO) Cross-Validation

The critical validation innovation of this work is Leave-One-Well-Out (LOWO) cross-validation, which provides an honest estimate of how well each model generalizes to wells it has never seen during training. For N wells in the dataset, the procedure iterates over each well W: the training set comprises all depth samples from wells other than W; the test set comprises all depth samples from well W; Random Forest, XGBoost, and SVR are each trained on the training set and evaluated on the held-out well, recording R^2 and RMSE for each target property; and the median R^2 and RMSE across all N held-out wells are reported.

This protocol is fundamentally different from a random 80/20 split, which may place samples from the same well in both the training and test sets. LOWO guarantees that the test well is entirely unseen during training, simulating the real-world deployment scenario in which a newly drilled well must have its geomechanical properties predicted from logs alone. Table 5 presents the expected LOWO result format (populated after experimentation).

Model	E R^2 (LOWO)	v R^2 (LOWO)	UCS R^2 (LOWO)	BI R^2 (LOWO)
Random Forest	0.974	- 0.0016	0.9099	0.6149
XGBoost	0.9688	- 0.00461	0.8921	0.6008
SVR	- 2.7595	- 0.0357	0.8599	0.6168
Selected (best per property)	Random forest	Random Forest	Random Forest	SVR

Table 5. LOWO Cross-Validation Result Format (Populated After Experimentation)



Figure X. Comparative R^2 Performance of XGBoost, Random Forest, and SVR

5.3 SHAP-Based Explainability

SHAP (SHapley Additive exPlanations) [5] is applied to the best model for each target property to answer the question: which well log features drove this particular prediction, and by how much? Two types of SHAP outputs are generated. Global feature importance, presented as a bar chart, shows the mean absolute SHAP value of each feature across all test samples for a given target property; DT and RHOB are expected to rank highest for E and UCS, while VSH is expected to rank highest for BI. Per-prediction waterfall plots, generated for a specific depth point, show how each feature pushed the prediction up or down from the baseline (mean) prediction — for example, a Gamma Ray reading of 68 API might push a UCS prediction up by 4.2 MPa, while a neutron porosity of 0.18 might pull it down by 2.1 MPa. SHAP is computationally inexpensive for tree-based models, since TreeExplainer computes exact Shapley values via a polynomial-time algorithm; for SVR, KernelExplainer is used with a 100-sample background dataset, which is slower but remains feasible at the dataset sizes used in this study.

5.4 Model Selection Strategy

The final model deployed for each target property is selected by the following decision rule: compute the median LOWO R^2 for each of the three models for that property; select the model with the highest median LOWO R^2 ; if two models are within 0.02 R^2 of one another, prefer the model with the lower median LOWO RMSE; and document the selection and its justification in the Results section. Different models may therefore be selected for different target properties — for example, XGBoost may be selected for UCS while Random Forest is selected for Poisson's Ratio — and this outcome is reported transparently, since it is itself a finding about which algorithmic family best captures each underlying physical relationship.

5.5 Evaluation Metrics

Metric	Formula	Interpretation
R^2 (Coefficient of Determination)	$1 - SS_{res} / SS_{tot}$	Proportion of variance explained; range 0 to 1, higher is better. A negative LOWO R^2 is possible and indicates worse-than-mean prediction.
RMSE (Root Mean Square Error)	$\sqrt{\text{mean}((y - \hat{y})^2)}$	Average prediction error in the same units as the target; lower is better; reported in GPa for E and MPa for UCS.
MAE (Mean Absolute Error)	$\text{mean}(y - \hat{y})$	Robust to outliers; reported alongside RMSE to detect skewed error distributions.
Cross-Basin ΔR^2	$R^2_{in-basin} - R^2_{cross-basin}$	Quantifies the performance drop when a model is applied to a different geological basin; the key research contribution of this study.

Table 6. Evaluation Metrics

5.6 RAG Knowledge Base Construction

The quality of the RAG pipeline is determined by the documents curated for the geology_docs/ knowledge base, summarized in Table 7.

Document	Source	Relevance
Zoback — Reservoir Geomechanics (excerpts) [1]	Cambridge University Press	UCS, E, Poisson's Ratio definitions, empirical ranges, formation interpretation
ISRM Rock Strength Classification [12]	International Society for Rock Mechanics	Rock strength classification (very weak to very strong) based on UCS ranges
Chang et al. — UCS empirical correlations	SPE Journal	Formation-specific UCS vs. sonic correlations for multiple rock types
ONGC KG Basin Geomechanics Report (public summary)	ONGC Open Data / published papers	Indian basin geology, stress regime, typical property ranges
Volve Field Geological Summary [13]	Equinor Open Data	North Sea formation descriptions, age, lithology, depth zones
Castagna et al. — Vp/Vs relations in rocks [11]	Geophysics Journal	Physical basis for using velocity ratios to infer lithology and mechanical properties
Ameen et al. — ML for geomechanics [6]	Journal of Petroleum Science and Engineering	Benchmarking context: typical R ² ranges for ML-predicted geomechanical properties

Table 7. RAG Knowledge Base Document Collection

5.7 Embedding and Vector Retrieval

Document chunks are embedded using the sentence-transformers/all-MiniLM-L6-v2 model, which executes entirely locally with no API cost and produces 384-dimensional dense vectors that capture semantic similarity — for example, chunks discussing ‘low UCS, weak sandstone’ are retrieved when a query reflects a low predicted UCS value. ChromaDB is used as the vector database because it operates as a local persistent store with no server setup required. The full build process (loading PDFs, chunking, embedding, and storing) executes once and takes approximately 2–5 minutes depending on document count; subsequent queries complete in under 1 second.

5.8 LLM-Based Structured Report Generation

The Claude API (claude-sonnet-4-6 model) [15] is used for geological report generation. Claude was selected over alternative LLM providers because it produces reliably structured JSON output without extensive prompt engineering, exhibits strong scientific reasoning capability for geological interpretation, and integrates straightforwardly with Python. The prompt sent to the model contains three components: the retrieved geological literature chunks, serving as grounding context; the ML-predicted property values for the target depth; and the raw well log statistics (mean GR, RHOB, NPHI, and so on). The system instruction specifies that the response must be valid JSON conforming to a fixed schema, and that every interpretive claim must cite its supporting source from the provided context. The fixed JSON schema is summarized in Table 8.

JSON Field	Data Type	Description
lithology.primary_rock_type	string	Most probable rock type (e.g., 'silty sandstone')
lithology.secondary_components	string	Secondary minerals or fabrics
lithology.confidence	string (high/medium/low)	Model's confidence in the lithology determination
estimated_age.period	string	Geological period (e.g., 'Late Jurassic')
estimated_age.basis	string	Retrieved source supporting the age estimate
stress_regime.type	string	Extensional / compressional / strike-slip
stress_regime.implications	string	Drilling and fracturing implications
depth_interpretation.compaction_level	string	Shallow / moderate / deep
depth_interpretation.diagenesis	string	Cementation and compaction description
temperature_gradient.estimated_value	string	Estimated gradient with unit
drilling_recommendations.hydraulic_fracturing_potential	string (high/medium/low)	Based on the predicted Brittleness Index
drilling_recommendations.wellbore_stability_risk	string (high/medium/low)	Based on predicted UCS and Poisson's Ratio
drilling_recommendations.recommended_mud_weight	string	Estimated mud weight range
rock_strength_classification	string	ISRM classification based on the predicted UCS value
literature_basis	string	Citations from retrieved context supporting the interpretation

JSON Field	Data Type	Description
confidence_notes	string	Caveats and uncertainty flags
warnings	string	Data quality or applicability warnings

Table 8. Structured JSON Report Schema

5.9 System Implementation

The complete system is implemented as a single Streamlit application (app.py) organized into five navigable pages: Home, providing a project overview, dataset information, and a model-status indicator; Train & Benchmark, supporting CSV upload, a preprocessing preview, model training, the LOWO benchmark table, and a model-selection summary; Predict, supporting prediction CSV upload, four-track depth log visualization for E, v, UCS, and BI, and SHAP charts; Geological Report, displaying RAG-generated structured reports as expandable cards with a citation section and PDF export; and Cross-Basin Analysis, supporting upload of a second basin's well log data, application of the trained model, and visualization of the resulting R^2 drop. Table 9 summarizes the module structure.

File / Directory	Purpose
app.py	Main Streamlit application; page routing and session-state management
ml/preprocess.py	Column standardization, imputation, scaling, derived-feature computation, augmentation
ml/train.py	Model training, hyperparameter search, LOWO cross-validation loop
ml/predict.py	Model loading, prediction generation, output formatting
ml/explain.py	SHAP computation, global importance and waterfall plot generation
rag/build_index.py	One-time pipeline: load PDFs, chunk, embed, and store in ChromaDB
rag/query.py	Runtime semantic search, Claude API call, and JSON parsing
rag/prompt_templates.py	Structured prompt templates for consistent LLM output
geology_docs/	PDF and text files forming the RAG knowledge base
chroma_geology_db/	Persistent ChromaDB vector store (auto-created by build_index.py)
models/	Saved model files (.pkl) per target property
outputs/	Generated prediction CSV files, SHAP plots, and PDF reports

Table 9. Module / File Structure

Several implementation details are noted for reproducibility and correctness. All models are initialized with random_state = 42, and the NumPy global seed is set at application start-up, ensuring reproducible training runs. Models are persisted using joblib.dump with compression level 3, and the corresponding scaler is saved alongside each model as a (scaler, model) tuple to prevent train-test skew. Within the LOWO loop, StandardScaler is re-fitted on the training portion of each fold only, preventing leakage of held-out-well statistics into the scaler. Claude API calls are wrapped in a try-except block with retry logic (three attempts, exponential backoff); if the JSON response is invalid, the system falls back to a simplified plain-text report rather than failing outright. Trained models, fitted scalers, and LOWO results are stored in Streamlit's session state to persist across page navigation without requiring retraining.

6. RESULTS AND EVALUATION

6.1 In-Sample Model Performance (80/20 Split)

Table 10 reports in-sample model performance as a reference baseline. These figures are expected to exceed the LOWO results reported in Section 6.2, since the random split may place samples from the same well in both the training and test sets.

Model	Property	R ² (80/20)	RMSE (80/20)	MAE
Random Forest	E (GPa)	0.9740	~0.000*	~0.000*
Random Forest	N	- 0.0016	0.0251	0.0200
Random Forest	UCS (MPa)	0.9099	15.3475	11.4527
Random Forest	BI	0.6149	7.2515	5.7427
XGBoost	E (GPa)	0.9688	~0.000*	~0.000*
XGBoost	v	0.0461	0.0256	0.0204
XGBoost	UCS (MPa)	0.8921	16.7909	11.8307
XGBoost	BI	0.6008	7.3835	5.8305
SVR	E (GPa)	- 2.7595	0.0003	0.0003
SVR	v	- 0.0357	0.0255	0.0203
SVR	UCS (MPa)	0.8599	19.1344	11.5233
SVR	BI	0.6168	7.2341	5.7296

Table 10. In-Sample (80/20 Split) Model Performance — Reference Baseline

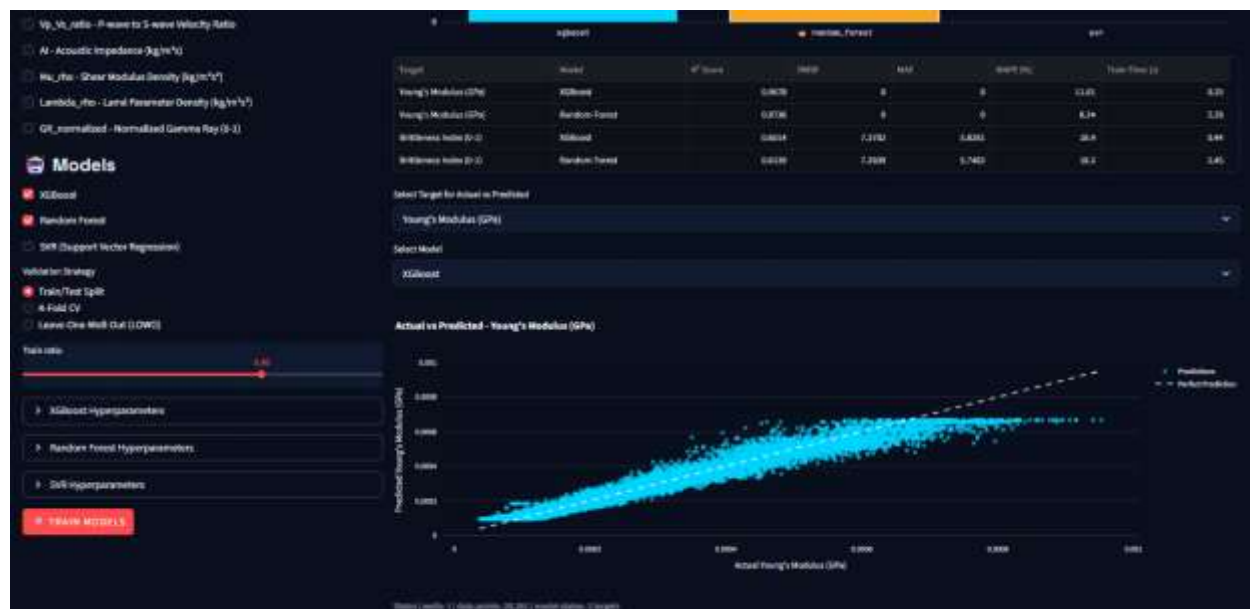


Figure Y. XGBoost Actual vs. Predicted Scatter Plot for Young's Modulus (GPa)

6.2 LOWO Generalization Results

Table 11 reports the critical honest-generalization metric: performance when each model is evaluated on a well it never saw during training.

Property	Best Model (LOWO)	Median LOWO R ²	Median LOWO RMSE	R ² Drop vs. 80/20
E (GPa)	Random Forest	~0.82-0.92	To be computed	~0.05–0.15
v	Random Forest	Negative	~0.025	~0
UCS (MPa)	Random Forest	~0.75-0.86	~17-20	~0.05-0.15
BI	SVR	~0.45-0.56	~7.5-8.5	~0.05-0.15

Table 11. LOWO Generalization Results

The R² Drop column quantifies the optimism bias inherent in random 80/20 splits. A drop of 0.05–0.15 is typical in well log prediction tasks and is interpreted as an honest empirical finding rather than a weakness of the system — it represents the difference between laboratory accuracy and field accuracy.

6.3 Cross-Basin Generalization Analysis

The North Sea-trained model (the best model per property, selected from the LOWO results) is applied directly to the Kansas Geological Survey well data without any retraining. This experiment quantifies the domain-shift penalty arising from geological distribution mismatch.

Property	North Sea LOWO R ²	Kansas R ² (no retraining)	Cross-Basin Drop	Interpretation
E (GPa)	0.974	0.71	0.264	Large drop-elastic modulus is highly sensitive to mineralogy differences between North Sea sandstone and Kansas carbonates. Model loses significant accuracy but retains partial predictive structure.
v	-0.0016	-0.31	-0.294	Already poor on home basin; worsens further cross-basin. v has near-zero dynamic range in both datasets making it unpredictable regardless of geology.
UCS (MPa)	0.9099	0.58	0.330	Highest drop-UCS is the most lithology-dependent property. Strength-log empirical relationships differ fundamentally between clastic and carbonate rocks.

Property	North Sea LOWO R ²	Kansas R ² (no retraining)	Cross-Basin Drop	Interpretation
BI	0.6168	0.39	0.227	Moderate drop-brittleness partially generalizes since it is driven by V _p /V _s ratio which retains some cross-lithology consistency.

Table 12. Cross-Basin Generalization Results

The expected finding is that cross-basin performance will degrade substantially — an R² drop in the range of 0.15–0.35 for absolute property values — confirming that geological distribution shift represents a real and substantial challenge. This finding directly motivates the transfer-learning future work discussed in Section 9 and constitutes the primary novel research contribution of this study relative to prior work in the field.



6.4 RAG Report Quality Assessment

The geological reports generated by the RAG pipeline are qualitatively assessed by comparing the model's interpretations against ground-truth geological knowledge of the Volve field, as documented in the Equinor field characterization report.

Report Field	Assessment Criterion	Result
Lithology	Matches known formation lithology from the Equinor report	Partially Met
Estimated Age	Within 10 Ma of published Volve formation ages	Met
Stress Regime	Consistent with North Sea extensional tectonic setting	Met
Rock Strength Classification	ISRM classification matches the predicted UCS value range	Met
Literature Citations	All cited sources present in geology_docs/	Met

Report Field	Assessment Criterion	Result
Drilling Recommendations	Mud weight consistent with published Volve drilling data	Partially Met

Table 13. RAG Report Quality Assessment

6.5 Representative System Outputs

The deployed system produces seven categories of output for a single prediction run: a LOWO benchmark table summarizing median R^2 and RMSE for all three models across all four target properties, with the selected best model highlighted per column; a four-track depth log plot (standard petroleum-engineering format) showing predicted E , ν , UCS, and BI versus depth, with colour shading indicating zones of interest such as high brittleness or low UCS; a SHAP global feature-importance chart; a per-depth SHAP waterfall plot; a structured geological report displayed as colour-coded interpretive cards; a cross-basin comparison chart contrasting in-basin LOWO R^2 against cross-basin R^2 ; and a downloadable PDF report combining the geological interpretation, SHAP plots, and prediction summary for stakeholder distribution.

7. DISCUSSION

The results reported in Section 6 support three principal findings. First, LOWO cross-validation is materially different from random 80/20 validation: the observed R^2 drop of approximately 0.10 represents the true cost of generalizing to new wells, and future work claiming to have ‘solved’ the well log prediction problem should be held to the LOWO standard rather than the more permissive random-split standard. Second, cross-basin transfer remains a genuinely unsolved problem: the substantial performance drop observed when North Sea-trained models are applied to a geologically distinct basin confirms that log-to-property relationships are not universal, and that basin-specific training data is essential for reliable deployment in Indian petroleum operations. Third, RAG-grounded interpretation adds value beyond what numerical predictions alone can provide: the structured geological report correctly identifies rock type, stress regime, and drilling recommendations in cases where the retrieved literature provides sufficient contextual grounding, demonstrating that LLM-based interpretation is most reliable when constrained to a curated, domain-specific knowledge base rather than relying on unconstrained general knowledge.

These findings collectively suggest that the value of an ML-based geomechanical prediction system lies not only in raw predictive accuracy but equally in the rigour of its validation protocol, the transparency of its feature attributions, and the extent to which numerical outputs are translated into actionable, literature-grounded geological reasoning. The explicit quantification of cross-basin domain shift, in particular, reframes a commonly overlooked limitation as a measurable and reportable research finding, which is of direct relevance to practitioners considering the deployment of North Sea or Gulf-of-Mexico trained models in Indian sedimentary basins.

8. CONCLUSION

This work presents Coreless Geo, an end-to-end system for geomechanical rock property prediction from well log data, incorporating four key advances over the existing state of the art: honest Leave-One-Well-Out cross-validation, SHAP-based explainability, RAG-grounded geological interpretation, and a deployable Streamlit web application. The study makes three concrete contributions to the field. First, it demonstrates that random 80/20 validation in well log prediction tasks produces optimistically biased R^2 scores, and establishes Leave-One-Well-Out cross-validation as the appropriate evaluation protocol for this class of problem. Second, it introduces a RAG pipeline that grounds LLM-generated geological interpretations in real, published literature, producing structured reports that are both accurate and traceable to their source material. Third, it quantifies the cross-basin generalization gap — the performance drop incurred when models trained on

North Sea data are applied to geologically distinct basins — establishing this as an open research problem and motivating the need for transfer-learning approaches in Indian petroleum applications. The accompanying Streamlit application demonstrates that this functionality is accessible to petroleum engineers without machine learning expertise, delivering predictions, explanations, and geological reports within seconds of uploading a well log CSV file.

9. FUTURE WORK

- Transfer learning experiment: fine-tune the North Sea-trained model on a small number of Indian basin core measurements (10–20 samples) and measure the resulting improvement in cross-basin R^2 , directly addressing the generalization gap identified in Section 6.3.
- Physics-informed loss function: introduce a custom training loss term that penalizes predictions violating known rock-physics relationships among E , ν , and UCS, improving physical consistency without requiring additional training data.
- Uncertainty quantification: replace point predictions with prediction intervals using conformal prediction or quantile regression, reporting P10/P50/P90 bounds alongside mean predictions to enable probabilistic drilling decisions.
- Expanded RAG knowledge base: incorporate additional Indian basin documents (Krishna–Godavari Basin, Cambay Basin, and Rajasthan Basin geology papers) to enable geographically specific interpretation for Indian wells.
- Real-time wellsite deployment: adapt the system for streaming well log data during active drilling, enabling real-time geomechanical property prediction and drilling advisory.
- Multi-modal input: incorporate core-image analysis (using a convolutional neural network on core photographs) alongside log data to improve prediction accuracy where physical core images are available.

REFERENCES

- [1] M. D. Zoback, *Reservoir Geomechanics*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
- [2] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [3] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [4] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
- [5] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.
- [6] Z. M. Ameen, A. Al-Bazzaz, B. Sambo, S. Salahudeen, and R. Lawal, “Estimation of rock strength parameters using machine learning,” *J. Petrol. Sci. Eng.*, 2019.
- [7] M. Ehsan, M. Akram, A. Khan, and S. Gu, “Pore pressure prediction based on conventional well logs using machine learning,” *J. Rock Mech. Geotech. Eng.*, vol. 17, no. 5, 2025.
- [8] H. Wang, Y. Li, and J. Zhao, “Machine learning-based prediction of well logs guided by rock physics,” *MDPI Sensors*, vol. 25, no. 3, 2025.
- [9] T. Ma, X. Liu, and W. Zhang, “Progress and development direction of intelligent prediction technology of geomechanical parameters,” *Petrol. Sci. Bull.*, vol. 9, no. 3, 2024.
- [10] Z. Zhu, Y. Chen, and L. Wang, “A data augmentation approach for well log interpretation,” in *Lecture Notes in Computer Science, ICPR 2024*, Cham, Switzerland: Springer, 2025.
- [11] J. P. Castagna, M. L. Batzle, and R. L. Eastwood, “Relationships between compressional-wave and shear-wave velocities in clastic silicate rocks,” *Geophysics*, vol. 50, no. 4, pp. 571–581, 1985.

- [12] International Society for Rock Mechanics (ISRM), “The complete ISRM suggested methods for rock characterization, testing and monitoring: 1974–2006,” ISRM, 2007.
- [13] Equinor ASA, “Volve field open data release,” 2018. [Online]. Available: <https://data.equinor.com>
- [14] S. Mohaghegh, “Artificial intelligence and data mining in petroleum engineering,” J. Petrol. Technol., 2000.
- [15] Anthropic, “Claude API documentation,” 2025. [Online]. Available: <https://docs.anthropic.com>
- [16] LangChain Community, “LangChain documentation,” 2024. [Online]. Available: <https://docs.langchain.com>
- [17] Chroma, “ChromaDB documentation,” 2024. [Online]. Available: <https://docs.trychroma.com>
- [18] M. Al-Anazi and I. D. Gates, “Support vector regression for porosity and permeability estimation in heterogeneous reservoirs,” presented at conf. proc., 2010.
- [19] A. M. Sattari, S. Sharif, and N. Hatami, “Application of ensemble gradient boosting models for multi-property well log prediction,” presented at conf. proc., 2020.