Rock Mineral Volume Inversion Using Statistical and Machine Learning Algorithms for Enhanced Geothermal Systems in Upper Rhine Graben, Eastern France

Pwavodi Joshua, Guy Marquis, Vincent Maurer, Carole Glaas, Anais Montagud, Jean-luc Formento, Albert Genter, Mathieu Darnet

JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000154

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

Rock Mineral Volume Inversion Using Statistical and Machine Learning Algorithms for Enhanced Geothermal Systems in Upper Rhine Graben, Eastern France

Pwavodi Joshua[1,2], Guy Marquis[1], Vincent Maurer[3], Carole Glaas[3], Anais Montagud[4], Jean-Luc Formento[4], Albert Genter[3], and Mathieu Darnet[2]

[1]Ecole et Observatoire des Sciences de la Terre, ITES UMR7063—CNRS/Université de Strasbourg, Strasbourg, France, [2]Bureau des Recherches Géologiques et Minières (BRGM), Orléans, France, [3]Electricité de Strasbourg—Géothermie, Mundolsheim, France, [4]CGG, Massy, France

Abstract Accurately determining the mineralogical composition of rocks is essential for precise assessments of key petrophysical properties like effective porosity, water saturation, clay volume, and permeability. Mineral volume inversion is particularly critical in geological contexts characterized by heterogeneity, such as in the Upper Rhine Graben (URG), where both carbonate and siliciclastic formations are prevalent. The estimation of mineral volumes poses challenges that involve both linear and nonlinear relationships associated with geophysical data. To address this complexity, our methodology strategically integrates the robust insights from standard statistical approaches with three machine learning (ML) algorithms: multi-layer perceptron, random forest regression, and gradient boosting regression. Furthermore, we propose a new hybrid ensemble model that incorporates a weighted average of multiple ML approaches to predict mineral composition within the Muschelkalk and Buntsandstein formations of the URG. ML techniques for mineral composition prediction in these formations exhibit robust predictive performance. The predicted mineral volumes align closely with quantitative estimates derived from X-ray diffraction analysis. Additionally, they are in good qualitative agreement with mineral descriptions obtained from cores and cuttings of the Muschelkalk and Buntsandstein formations.

Plain Language Summary We conducted an assessment of subsurface rock mineral compositions from their physical properties measured through logging tools, employing a combination of statistical and machine learning techniques. The outcomes derived from these methodologies demonstrate their complementary nature and robustness in elucidating the spatial distribution of minerals within Triassic rocks from the Upper Rhine Graben in France. This approach helps in deciphering complex mineralogical compositions and geological structures within subsurface geothermal reservoirs.

Key Points:

- We use well-logs to estimate mineral volumes in Triassic formations of the Upper Rhine Graben using statistical and machine learning (ML) methods
- We show that the random forest and gradient-boosting regressions provide better-fitting predictions than the multi-layer perceptron
- More realistic mineral volume estimates can be obtained by combining several ML algorithms rather than using a single one

Correspondence to:
P. Joshua,
j.pwavodi@brgm.fr

Citation:
Joshua, P., Marquis, G., Maurer, V., Glaas, C., Montagud, A., Formento, J.-L., et al. (2024). Rock mineral volume inversion using statistical and machine learning algorithms for enhanced geothermal systems in Upper Rhine Graben, eastern France. *Journal of Geophysical Research: Machine Learning and Computation*, *1*, e2024JH000154. https://doi.org/10.1029/2024JH000154

Received 30 JAN 2024
Accepted 3 JUN 2024
Corrected 10 JUL 2024

This article was corrected on 10 JUL 2024. See the end of the full text for details.

1. Introduction

Electricity and heat generation through enhanced geothermal system (EGS) technology hinge upon our ability to understand and predict the characteristics of hydrothermal fluids and rock formations (Darnet et al., 2023). Many conventional methods for assessing these properties are not cost-effective, underscoring the significance of relying on predictive techniques that can offer valuable insights during the early exploration stages. In the initial stages of comprehending EGS, assessing petrophysical properties plays a crucial role. To fully evaluate the petrophysical properties of EGS reservoirs or formations, it is very important to know as accurately as possible the mineralogical composition for accurate estimation of rocks' effective porosity, water saturation, clay volume, and permeability (Hosseini, 2018; Zhao et al., 2016).

The most comprehensive effort to characterize the mineralogical composition of the various formations within the URG Basin was an extensive synthesis of geological outcrop samples, wireline gamma ray (GR) logs and core descriptions from deep wells such as EPS-1 drilled in the 1990s (Aichholzer et al., 2019; Duringer et al., 2019). Among the major formations identified in their work, the Buntsandstein formation, predominantly composed of sandstone with the presence of clays, and the Muschelkalk formation, characterized by calcite, dolomite, clays, and anhydrite minerals (Aichholzer et al., 2016, 2019; Duringer et al., 2019). While the research by Aichholzer et al. (2019) and Duringer et al. (2019) successfully delineated the primary facies descriptions of these formations

Author Contributions:
Conceptualization: Pwavodi Joshua
Data curation: Pwavodi Joshua, Guy Marquis, Vincent Maurer, Carole Glaas, Anais Montagud, Jean-Luc Formento, Albert Genter
Formal analysis: Pwavodi Joshua, Anais Montagud
Funding acquisition: Guy Marquis, Jean-Luc Formento, Albert Genter, Mathieu Darnet
Investigation: Pwavodi Joshua, Anais Montagud, Jean-Luc Formento
Methodology: Pwavodi Joshua, Anais Montagud, Jean-Luc Formento
Project administration: Guy Marquis, Jean-Luc Formento, Albert Genter, Mathieu Darnet
Resources: Pwavodi Joshua, Guy Marquis, Vincent Maurer, Carole Glaas
Software: Pwavodi Joshua, Anais Montagud, Jean-Luc Formento
Supervision: Guy Marquis, Vincent Maurer, Jean-Luc Formento, Albert Genter, Mathieu Darnet
Validation: Pwavodi Joshua, Guy Marquis, Carole Glaas
Visualization: Pwavodi Joshua
Writing – original draft: Pwavodi Joshua
Writing – review & editing: Pwavodi Joshua, Guy Marquis, Vincent Maurer, Carole Glaas, Anais Montagud, Jean-Luc Formento, Albert Genter, Mathieu Darnet

and identified their major mineral constituents, a precise quantitative estimation of the volume of each mineral constituent within the siliciclastic and carbonate formations is yet to be achieved.

There are several different approaches described in the literature for analyzing mineral compositions. Among them, X-ray diffraction (XRD) mineral analysis is recognized as the most precise means of mineral estimation (Eberl, 2003; Heap et al., 2017, 2019; Hillier, 1999; Xiao et al., 2023). In the URG, limited quantitative bulk mineral estimates from XRD analysis of samples from the Middle—Lower Muschelkalk and Buntsandstein along EPS-1 well were published by Heap et al. (2017) and Heap et al. (2019). Nevertheless, when it comes to well-scale applications, this approach does have its limitations, primarily stemming from challenges associated with the unavailability of core samples along the entire well length, which are often constrained by cost considerations (Pwavodi & Doan, 2023). Additionally, XRD analysis on borehole cutting samples is feasible; however, precision in determining lithological intervals may be compromised. In fact, in certain drilling methods, like the riserless drilling system, utilizing cuttings is not possible as they are completely lost to the seafloor and not returned to the driller at the surface through riser pipes (Pwavodi, 2023). Consequently, characterizing the distribution of minerals along the entire length of a well using this method can prove to be a challenging task. According to the history of drilling for EGS in the Soultz-sous-Forêts and Rittershoffen sites of the URG, only the EPS-1 well was continuously cored along its length (Aichholzer et al., 2019; Genter & Traineau, 1991, 1996). Consequently, alternative methods for estimating mineralogy in wells with limited or no core samples within the URG have become necessary.

Savre (1963) proposed a simple compositional mathematical solution model to estimate the proportions of minerals in rocks using a triangular coordinate graph approach. This method is further solved within an iterative inversion linear-log system, in which the well-log response is calculated as the sum of the product between the multiple minerals' (e.g., quartz, calcite, clay, dolomite, etc.) volumes and their average theoretical response values (Amosu & Sun, 2018; Doveton, 2014). This linear system is written for different log responses such as bulk density, photoelectric factor, acoustic slowness, or neutron porosity (Doveton, 2014). All the linear equations for the different log responses are solved with a matrix inversion system with the vector of the unknown proportions being estimated. This method looks indeed simple to implement, but several constraints need to be added to the equation system to avoid resulting in negative volume values coming from the matrix inversion process, instrumental errors, or poor well environment. Some of the well logs like the sonic compressional velocity cannot be determined with linear systems, hence, a solution that factors in the nonlinearity of logs is needed during the inversion.

Despite the effectiveness of these mineral volume estimation methods, there is a continuing need for a more predictive and accurate approach to estimate mineral volumes in different stratigraphic lithologies along wells, especially in the absence of cores, cuttings, or wireline logs. This approach should not be limited by prior knowledge of theoretical individual mineral response values, tool errors, well conditions, optimization problems, or other unknown constraints that may affect the accurate estimation of mineral volumes. In response to these challenges, advances in artificial intelligence (AI) have been leveraged. In recent years, several investigations have delved into diverse machine learning (ML) methodologies for mineral composition estimation. Hu et al. (2023) devised a hybrid ML framework, amalgamating convolutional neural network architecture with XGBoost, while Laalam et al. (2022) conducted a comparative analysis of the performance of linear regression (LR), support vector regression, random forest regression (RFR), extra trees regression, K-nearest neighbors, and extreme gradient booster (XGBoost). Conversely, Lee and Lumley (2023) assessed the mineralogical brittleness index of shaly formations employing a blend of statistical and ML techniques, including decision trees, ensembles, support vector machines, probabilistic neural network, and deep feedforward neural network. Deng et al. (2019) proposed an optimized Bayesian inversion approach for estimating rock petrophysical and compositional properties. Mustafa et al. (2022) estimated shale mineralogy and elastic properties using the adaptive neuro-fuzzy inference system and artificial neural networks.

However, there exists a significant opportunity to deepen our comprehension of the mineralogical composition of both carbonate and siliciclastic sediments. Thus, in this study, we utilized four distinct methodologies to address this knowledge gap. Specifically, we employed RFR, multi-layer perceptron (MLP), gradient boosting regression (GBR), and a hybrid ensemble approach. The choice of these algorithms for this work was driven by their adaptability, scalability, and robustness in handling complex data sets and addressing nonlinear problems (Bishop, 1995; Chen & Guestrin, 2016; Liaw & Wiener, 2002) such as mineral volume inversion. We show here

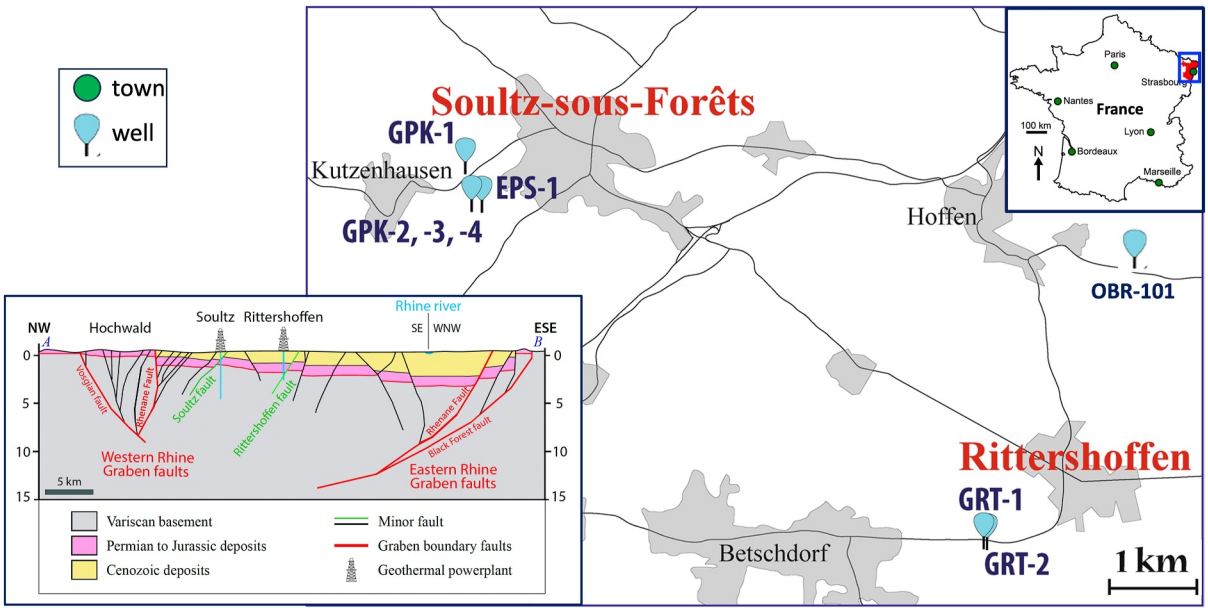

Figure 1. Summary of the study area: at the top right corner of the picture presents the Map of France showing the location of the Bas-Rhin (Lower Rhine) department of Alsace (in red) (modified from Michael et al. (2019)). It also captures the locations of the heads of the wells discussed in this paper. These positions are within the northwestern region of the Upper Rhine Graben (URG) (modified from Aichholzer et al. (2019)). The bottom left inset shows a schematic geological cross-section through the URG at the latitude of Rittershoffen and Soultz-sous-Forêts (adapted from Brun et al. (1992), Kappelmeyer et al. (1991)).

that these AI-based techniques represent a promising means to enhance the accuracy and the reliability of mineral volume estimations in geological settings.

2. Geology of the Upper Rhine Graben, NE France

The URG was formed during the Late Eocene in response to the NNE-trending Alpine compression (Illies, 1972; Rotstein et al., 2006; Villemin & Bergerat, 1987; Villemin et al., 1986). It is a prominent complex asymmetrical extensional graben approximately 300 km long and 30 km wide, on both the French and German sides of the Rhine River (Rotstein et al., 2006). It is delineated by both eastern and western systems of major faults (Figure 1), which separate the sediment-filled graben (with up to 3.5 km of tertiary sediments) from the uplifted graben shoulders (Vosges on the French and Black Forest on the German sides, respectively) (Duringer et al., 2019). Additionally, the URG comprises several sub-basins each with its own features (Duringer et al., 2019).

The URG has been identified as a major target for deep geothermal exploration thanks to the high geothermal gradient that can locally reach more than 100°C/km, for example, under Soultz-sous-Forêts (Agemar et al., 2012; Baillieux et al., 2013; Pribnow & Schellschmidt, 2000), and fault zones acting as potential fluid pathways (Bächler et al., 2003; Duwiquet et al., 2021; Guillou-Frottier et al., 2013). For over 30 years, it has been the object of deep drilling to explore for geothermal resources, with a particular focus on the sites of Soultz-sous-Forêts and Rittersoffen (Figure 1) (Duringer et al., 2019). While several wells have been drilled in these areas, our study centers on four wells: EPS-1 and GPK-1 in Soultz-sous-Forêts, and GRT-1 and OBR-101 in Rittershoffen and Oberroedern, as depicted in Figure 1 (Aichholzer et al., 2016, 2019; Genter & Traineau, 1992).

In Soultz-sous-Forêts, deep geothermal drilling started in 1987, with GPK-1 as the first geothermal borehole initially drilled to a depth of 2,000 m, intersecting the top of the granitic basement at the depth of 1,376 m (Genter & Traineau, 1992, 1996). Subsequently, in 1992, it was further extended to a depth of 3,600 m (Genter & Traineau, 1992, 1996). Nearby well EPS-1 was originally drilled for hydrocarbon exploration, known by the designation No. 4589. It was however extended and cored to the granitic section at a depth of about 2,227 m in 1990–1991 (Genter & Traineau, 1992).

GRT-1 is the first geothermal borehole that was drilled in Rittershoffen in 2012. It was drilled to a depth of 2,580 m under the fractured granitic basement rock, and it is reported to have intersected the Rittershoffen fault at

Table 1
Coupling Parameters Used in Modeling

Log response/minerals	Calcite	Clay	Dolomite	Anhydrite	Quartz
ρ (kgm^{-3})	2.710	2.680	2.870	2.950	2.650
P_e (barns cm^{-3})	5.090	3.030	3.13	5.080	1.810
K (gAPI)	60.8	32.8	38.8	54.9	32.8
V_p (km/s)	6.457	4.740	6.943	6.096	4.689

a depth of 2,400 m (Baujard et al., 2017). The Oberroedern well (OBR-101) produced hydrocarbons throughout the 1970s and 1980s and also produced significant amounts of hot salty water (Munck et al., 1979).

Critical to comprehending EGSs is the fact that the four wells in this study intersect major lithologies at varying depths in the regional geologic sequence. As evidenced by Baujard et al. (2017), Aichholzer et al. (2016, 2019), and Duringer et al. (2019) these lithologies include Triassic sediments (Muschelkalk limestone and Buntsandstein sandstones), Permian clastic sandstones, and a Paleozoic crystalline basement composed of hydrothermally altered and fractured granite within fresh granite.

3. Methodology

As stated above, our study focuses on four wells: OBR-101 and GRT-1 near Rittershoffen and EPS-1 and GPK-1 near Soultz-sous-Forêts. We adopted a two-pronged approach to estimate mineral volumes

1. We employed a statistical method that centers on solving linear systems. This approach enables us to calculate mineral volumes based on well-log data and established relationships, particularly suited for well conditions where linear assumptions are valid.
2. We leveraged the power of machine-learning algorithms. These algorithms excel at solving nonlinear systems, offering a more flexible and data-driven alternative for estimating mineral volumes.

By integrating both linear and nonlinear methods, we aim to provide a comprehensive and accurate assessment of mineral volumes at these well locations.

3.1. Mineral Volume Inversion Using Statistical Approach

We first solve the mineral volume using an industry software which is based on the statistical methodology introduced by Mitchell and Nelson (1988). Its core functionality is solving linear systems, making it well-suited for mineral volume models, whether they are balanced, over-determined, or under-determined (Mitchell & Nelson, 1988). The inversion process seeks to minimize the misfit between the model response to a set of volume estimates and the observed normalized and reconstructed theoretical tool responses (Mitchell & Nelson, 1988). The misfit estimation function Δ to minimize is as follows:

$$\Delta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(m_i - f_i)^2}{\tau_i^2 + \delta_i^2}} \tag{1}$$

where n is the number of tool response equations, m_i is the actual measurement of tool i, f_i is the theoretical response of tool i (calculated from theoretical tool response equation), τ_i^2 is the variance of the error of the measurement of tool i, and δ_i^2 is the variance of the error of the theoretical response of tool i. The theoretical response of tool i is computed using the following equation.

$$f_i = \sum_{j=1}^{m} e_{ij} V_j \tag{2}$$

where V_j is the fractional volume of mineral j, m is the number of minerals of interest, e_{ij} is the mineral endpoint of tool i in mineral j (Table 1). To solve theoretical tool responses, as outlined in Equation 1, within a simple mineral volume model, we estimate the proportions of various components in conjunction with porosity. This estimation is derived from a system of equations based on the measured log responses. These linear equations establish a connection between the log measurements and the properties of the mineral and fluid constituents:

$$CV_j = L \tag{3}$$

where C is a matrix of the component petrophysical properties, and L is a vector of the well-log responses over the evaluated zone, which represent the bulk petrophysical properties of the rock formation. In this study, the primary formations under consideration are the Muschelkalk (carbonate) and Buntsandstein (siliciclastic) formations. As an initial step, the main mineral constituents of the Muschelkalk formations (calcite, dolomite, clay, anhydrite, and quartz) are identified based on evidence obtained from core and field outcrop samples, as documented by Aichholzer et al. (2016, 2019), and Duringer et al. (2019). For multiple mineral systems, Equation 3 can be adapted for the different Muschelkalk minerals and solved using a linear system:

$$\begin{bmatrix} \rho_{calc} & \rho_{Clay} & \rho_{dol} & \rho_{anh} & \rho_{qtz} \\ Pe_{calc} & Pe_{Clay} & Pe_{dol} & Pe_{anh} & Pe_{qtz} \\ K_{calc} & K_{Clay} & K_{dol} & K_{anh} & K_{qtz} \\ Vp_{calc} & Vp_{Clay} & Vp_{dol} & Vp_{anh} & Vp_{qtz} \end{bmatrix} \begin{bmatrix} V_{calc} \\ V_{Clay} \\ V_{dol} \\ V_{anh} \\ V_{qtz} \end{bmatrix} = \begin{bmatrix} \rho_{\log} \\ Pe_{\log} \\ K_{\log} \\ Vp_{\log} \end{bmatrix} \tag{4}$$

The different symbols in Equation 4 are explained in Table 2. The system of Equation 4 can be easily solved (Penrose, 1955). To compare the theoretical response with the actual tool measurements (Figure 2), as described in Equation 1, two constraints were added to ensure physically meaningful results: (a) the unity constraint, that is, the sum of all mineral volumes must equal one and (b) the positivity constraint, that is, no individual mineral volume can be allowed to be less than zero. For well-logs exhibiting nonlinear behavior, such as VP, PEF, and NPHI, a linearization step was necessary before implementing the equations. The Muschelkalk depth interval was divided into two parts for the equation: the upper part considered calcite, dolomite, clay, and anhydrite, with the quartz volume initialized to zero. In the lower depth interval, only three minerals were considered: quartz, dolomite, and clays with the volumes of calcite and anhydrite initialized to zero. This division was applied to estimate mineral volumes for both the OBR-101 and GRT-1 wells. To handle these divisions within the equation, depth constraints were incorporated into a loop workflow (Figure 2), facilitating the step-by-step solution of the mineral volumes.

3.2. Mineral Volume Inversion Using Artificial Intelligence

Artificial neural networks are inspired by the structural and functional aspects of the human nervous system (Aggarwal, 2018; McCulloch & Pitts, 1943; Rosenblatt, 1958). These networks comprise interconnected nodes, or neurons, linked by synaptic connections. In our study, we employed the MLP, the RFR, and the GBR to model and estimate the volume of minerals within the Muschelkalk and Buntsandstein formations (Figure 3). Additionally, we introduced a hybrid ensemble method that combines the results from MLP, GBR, and RFR using a weighted averaging technique. The models were trained and tested using a data set consisting of well-log data, including parameters such as sonic travel-time (DT), photoelectric factor (PEF), and GR gathered from three distinct wells. In the initial stages of neural network modeling (Figure 3), a critical data cleaning process was conducted to ensure the quality and reliability of the data set.

3.2.1. Data Preprocessing

Data exploration and preprocessing were done to remove unwanted data points (mostly negative values and outliers) that could increase the bias of the model or its error bar. Simple moving average (SMA) was used to reduce data scatter. It calculates the unweighted mean of the preceding k data points. The higher the value of k, the smoother the curve, but less accurate the result. We compute the SMA if the data points are p_1, p_2, …, p_n:

$$SMA_k = \frac{1}{k} \sum_{i=n-k+1}^{n} p_i \tag{5}$$

In addition to SMA, we performed log conditioning, editing, depth shifts, and used computer vision technique to extract legacy data that were only available on paper.

Table 2
List of Symbols and Notations

Symbol or acronym	Meaning
ρ_{calc}	Matrix density of calcite in kg/m^3
ρ_{clay}	Matrix density of clay in kg/m^3
ρ_{dol}	Matrix density of dolomite in kg/m^3
$\rho_{anhydrite}$	Matrix density of calcite in kg/m^3
ρ_{qtz}	Matrix density of quartz in kg/m^3
ϕ	Porosity of the rock formation (pu)
Pe_{calc}	Photoelectric factor of the calcite matrix
Pe_{clay}	Photoelectric factor of the clay matrix
Pe_{dol}	Photoelectric factor of the dolomite matrix
$Pe_{anhydrite}$	Photoelectric factor of the calcite matrix
Pe_{qtz}	Photoelectric factor of the quartz matrix
K_{calc}	Spectral potassium log value of the calcite matrix
K_{clay}	Spectral potassium log value of the clay matrix
K_{dol}	Spectral potassium log value of the dolomite matrix
$K_{anhydrite}$	Spectral potassium log value of the calcite matrix
K_{qtz}	Spectral potassium log value of the quartz matrix
Vp_{calc}	Sonic velocity response of the calcite matrix
Vp_{clay}	Sonic velocity response of the clay matrix
Vp_{dol}	Sonic velocity response of the dolomite matrix
$Vp_{anhydrite}$	Sonic velocity response of the calcite matrix
Vp_{qtz}	Sonic velocity response of the quartz matrix
V_{calc}	Calcite volume
V_{clay}	Clay volume
V_{dol}	Dolomite volume
$V_{anhydrite}$	Anhydrite volume
V_{qtz}	Quartz volume
$\rho t_{\log}$	Bulk density measured well log
$Pe_{\log}$	Photoelectric factor measured well log
$\Delta t_{\log}$	Sonic transit measured well log
$Vp_{\log}$	Sonic velocity measured well log

3.2.2. Feature Selection

We initially used the Pearson correlation coefficient to assess the relationship between input variables and identify multicollinearity. This metric assesses the strength and direction of the linear relationship between two continuous variables. Pearson's correlation coefficient, denoted as $P_{x,y}$, ranges from -1 (indicating a perfect negative linear relationship) to $+1$ (representing a perfect positive linear relationship). Mathematically, $P_{x,y}$ is expressed as follows:

$$P_{x,y} = \frac{\mathrm{Cov}(x,y)}{dx\,dy} \tag{6}$$

where Cov represents the covariance, and dx and dy are the standard deviations of input variables x and y, respectively. Upon analyzing the correlation coefficients (Figure 4), we observed that the variable DT exhibited a notably high negative correlation with Vp, with a Pearson correlation coefficient of approximately -0.95. Ideally, such high correlation might suggest multicollinearity issues, prompting consideration for dropping one of the variables to mitigate redundancy. However, in the context of data augmentation, we opted to retain both DT and Vp as part of the input data set. This decision was influenced by the lack of similar data sets across all wells used in our study, highlighting the importance of maintaining comprehensive input variables despite potential multicollinearity concerns.

Feature selection was executed using a stepwise bidirectional approach to ascertain the individual contribution of each well-log variable in enhancing the efficacy of machine-learning model training (Yu & Liu, 2003). This approach systematically assesses the statistical significance of each independent variable within a LR model (Siddiqi et al., 2022; Yu & Liu, 2003). It harnesses the dual advantages of both forward selection and backward regression elimination techniques to discard undesirable features (Siddiqi et al., 2022; Yu & Liu, 2003). The forward regression method commences by identifying the most influential well-log variables, such as DT, and progressively augments the overall set of well-logs (as illustrated in Figure 5a). The criterion for determining the order of inclusion of these variables is the R-squared (R^2) score, a metric that profoundly influences the process of introducing new entries and the ultimate selection of values:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{7}$$

where $\hat{y}_i$ is the predicted well-log data, that is, a well-log data value on the regression line, $\bar{y}_i$ is the mean of the well-log, y is the well-log data, SST is the sum of squares, and SSR is the sum of squares of residuals.

The backward regression is utilized to carry out deletion, which is also referred to as backward elimination (Figure 5b). If the R^2-score value that is being tested is the lowest possible, then the well-log with the lowest R^2-score is eliminated from the stack. The results of the forward selection and backward regression elimination techniques were similar on the basis of their degree of confidence based on their evaluation criterion (Figure 6 and Table 3).

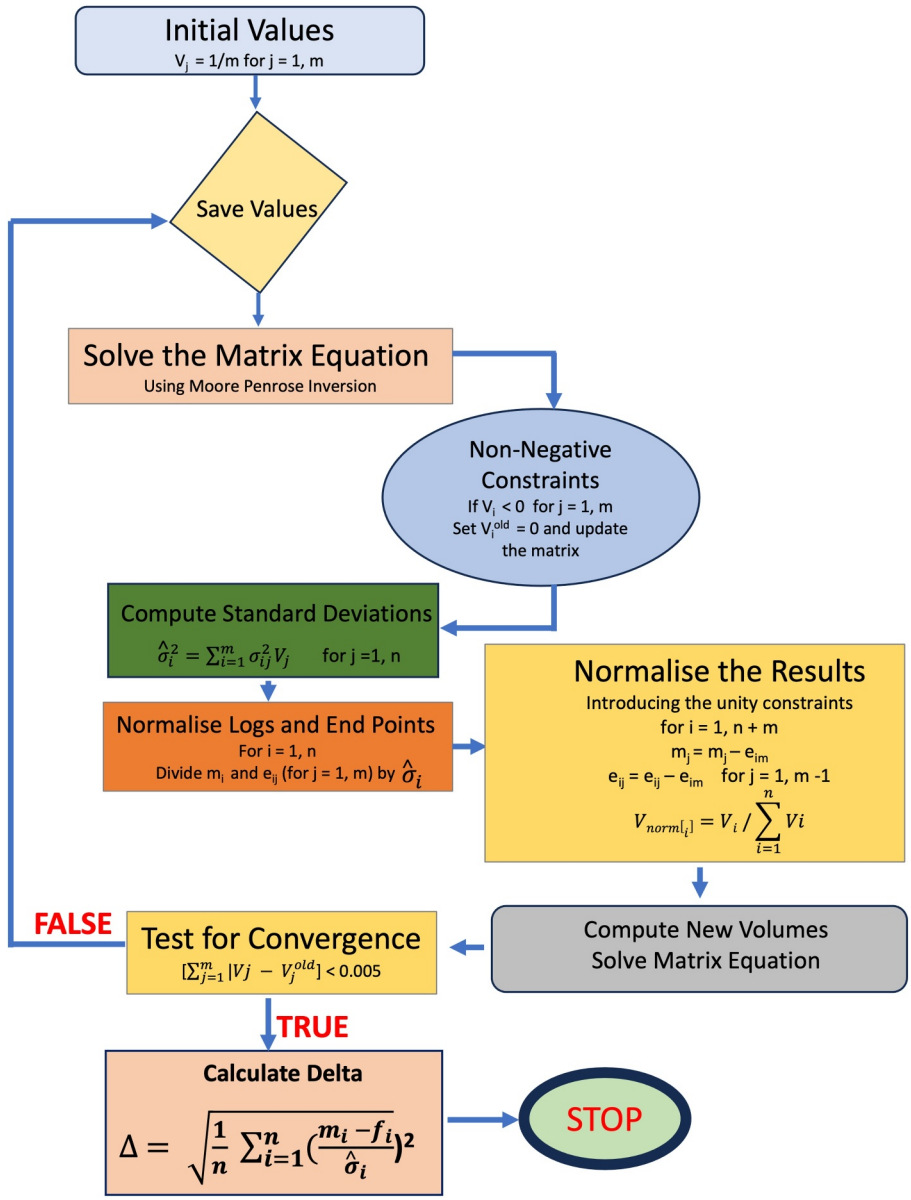

Figure 2. The workflow describing the statistical approach (modified from Mitchell and Nelson (1988)).

3.2.3. Feature Scaling and Data Splitting

The data set was divided into distinct categories: independent and dependent variables. The independent variables constitute a quintet of well-log data sets, while the dependent variables are the mineral volumes to be ascertained through the statistical approach. The data sets were further constrained within a finite range using the standardization technique: we computed the *Z*-score (Z_s) for each data point, that is, all attributes were centered with respect to a mean value of zero and a standard deviation of one.

$$Z_s = \frac{X_{\log} - \mu_{\log}}{\sigma} \tag{8}$$

where $X_{\log}$ is the well log value, $\mu_{\log}$ is the mean value, and σ is the standard deviation. Following the transformation of both independent and dependent variables, the data set was split into *train*, *validation*, and *test* data sets. In this context, we consider the data set to be partitioned according to a specified proportion denoted as *p*,

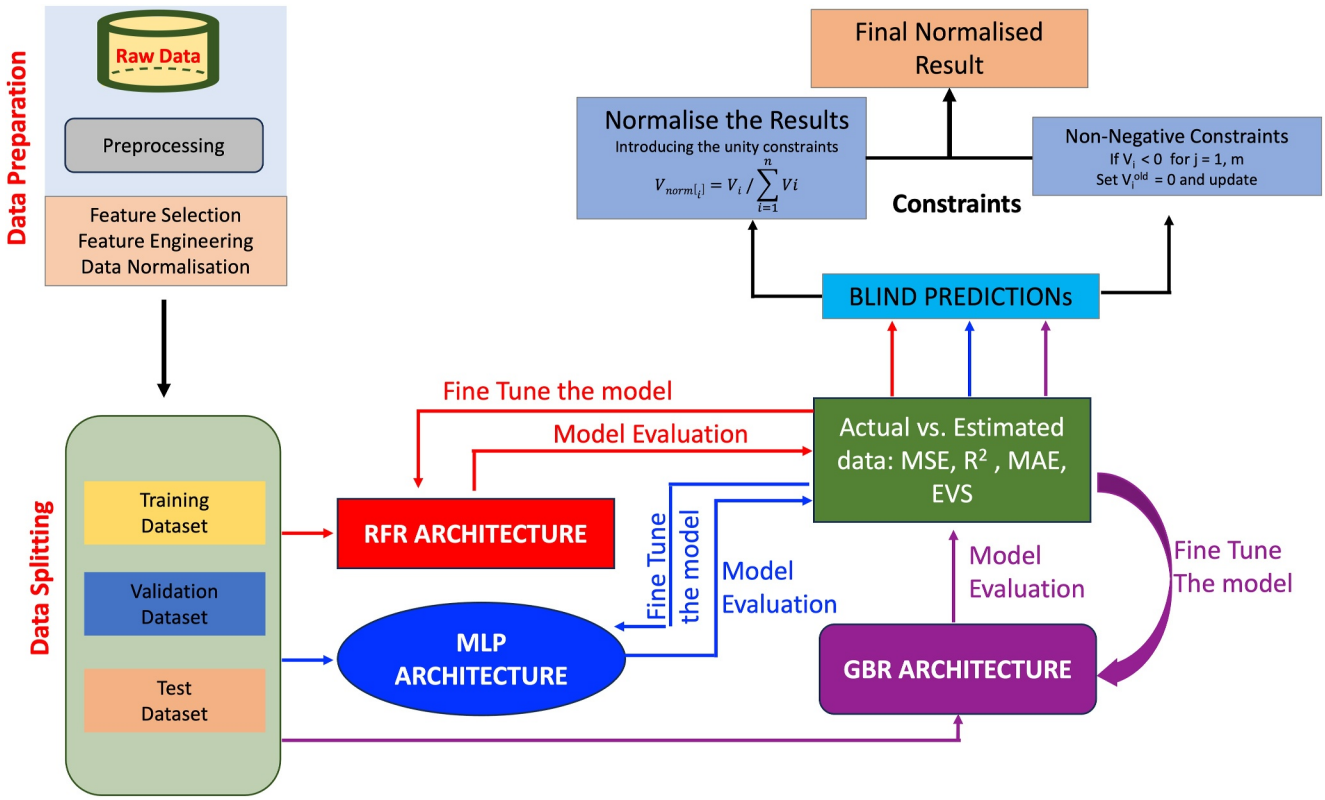

Figure 3. The overview of the Machine Learning Workflow used for mineral volume estimation in this study.

typically confined within the interval $0 < p < 1$. This data set splitting results in the creation of two index sets, namely $I_{training}$ and $I_{testing}$. The size of $|I_{training}| = [p.n]$ and $|I_{testing}| = n - |I_{train}|$, where n is the data set's total size. In this study, the neural network methodologies were trained using 80% of the data set, while the remaining 20% was allocated for testing. Additionally, 10% of the training data set was dedicated for validation purposes.

3.2.4. Model Evaluation Metrics

All the outcomes generated by the ML models will be evaluated by four fundamental assessment metrics: mean squared error (MSE), mean absolute error (MAE), R-squared (R2-score), and explained variance score (EVS). The MSE quantifies the average of the squared differences between predicted and actual values. This metric is particularly advantageous in regression tasks as it assists in fine-tuning models and precisely assesses the magnitude of errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{9}$$

where y_i and $\hat{y}_i$ are, respectively, the measured and predicted well-log value i. In contrast, the MAE measures the average of the absolute differences between predicted and actual values, offering greater resilience to outliers.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

The R^2 score (Equation 7) reveals the proportion of the variance of the dependent variable that can be explained using independent variables. *EVS* shares similarities with R^2 and scores close to 1.0 are optimal. *EVS* complements R^2 by providing insights into the predictive power of the model concerning the variance in the data

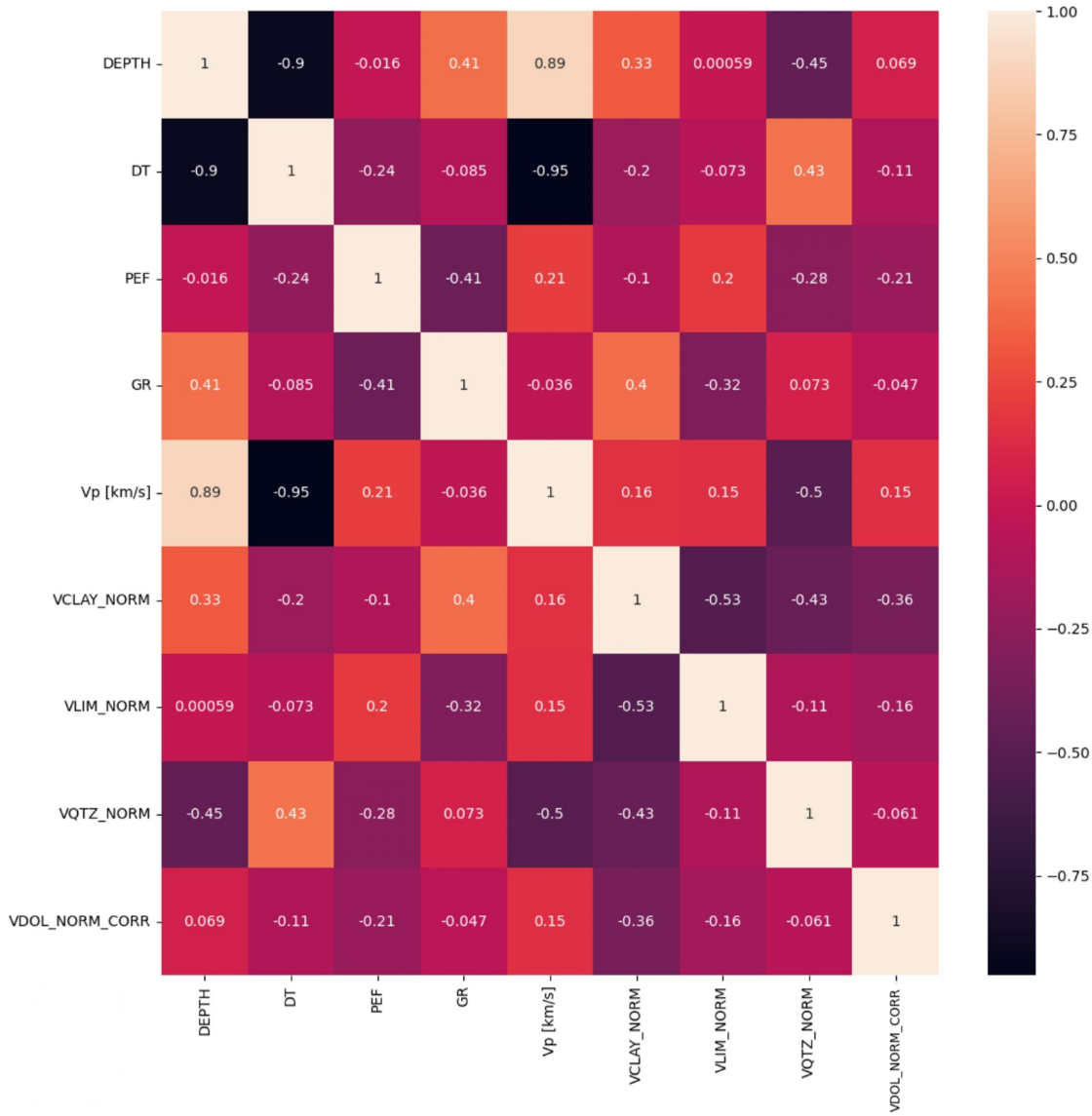

Figure 4. Pearson correlation coefficients of input variables.

$$EVS = 1 - \frac{\mathrm{Var}(y_i - \hat{y}_i)}{\mathrm{Var}(y_i)} \tag{11}$$

3.2.5. Mineral Volume Inversion Using Multi-Layer Perception (MLP)

MLP is a supervised learning algorithm that learns a function $f(\cdot)$: $R^n \rightarrow R^y$ by training on a data set, where n is the number of dimensions for input variables and y is the number of dimensions for output variables. MLP was first proposed by Rosenblatt (1958) and later enhanced by incorporating nonlinearity through the utilization of stochastic gradient descent for the purpose of classifying patterns (Amari, 1967; Ivakhnenko, 1967). Moreover, the method underwent significant enhancements, resulting in the development of a backpropagation method. This advancement was achieved through the incorporation of a supervised learning strategy based on the chain rule (Rodriguez & Lopez Fernandez, 2010).

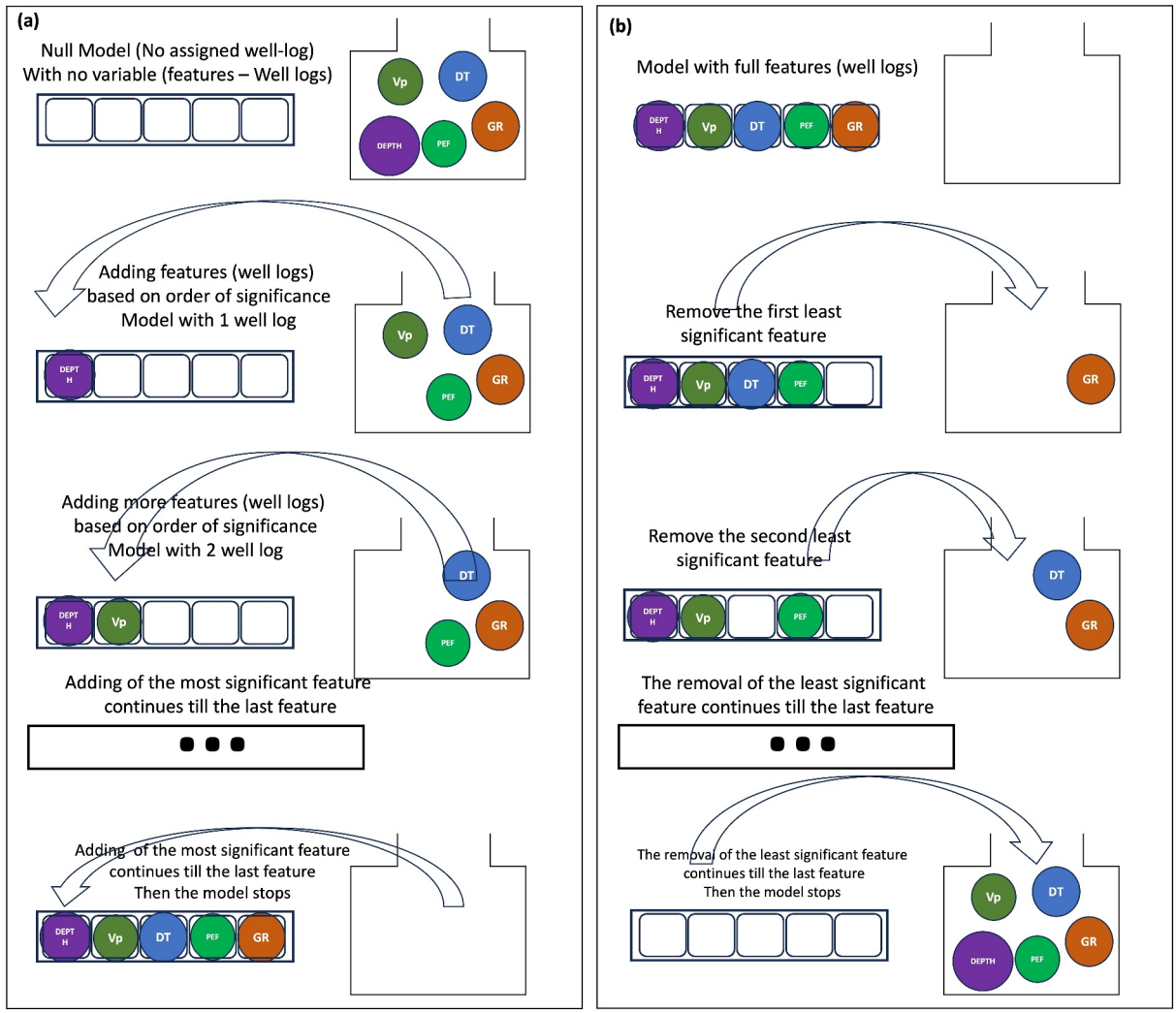

Figure 5. A schematic diagram showing the (a) stepwise forward regression and backward selection method used for well logs selection.

Leveraging the intrinsic capacity of MLP to tackle nonlinearity issues (Amari, 1967; Ivakhnenko, 1967; Rodriguez & Lopez Fernandez, 2010), this study employed MLP to address the inherent nonlinear behaviors observed in input well logs, such as sonic velocity (Vp), PEF and neutron porosity (NPHI). Unlike the statistical approach discussed earlier, MLP can be effectively employed without the necessity for prior linearization of the well logs. In this context, the set of well logs previously categorized as independent variables, denoted as $X = x_1$, x_2, $\ldots x_n$, while the independent variables (mineral volumes) as the target (y) and the bias unit (b) were used. The target (y) for a simple multi-layered neural network can be estimated as the sum of the product of weights (w_i) and input well-logs (x_i) augmented by the bias term (b) (Rodriguez & Lopez Fernandez, 2010). This can be formally expressed as follows:

$$y = \sum_{i=1}^{m} w_i x_i + b \tag{12}$$

To introduce nonlinearity into Equation 12, differentiating ML techniques from the multiple linear algebra systems employed for mineral volume inversion (as illustrated in Equation 4), a nonlinear function, commonly known as the activation function (denoted as σ), was introduced:

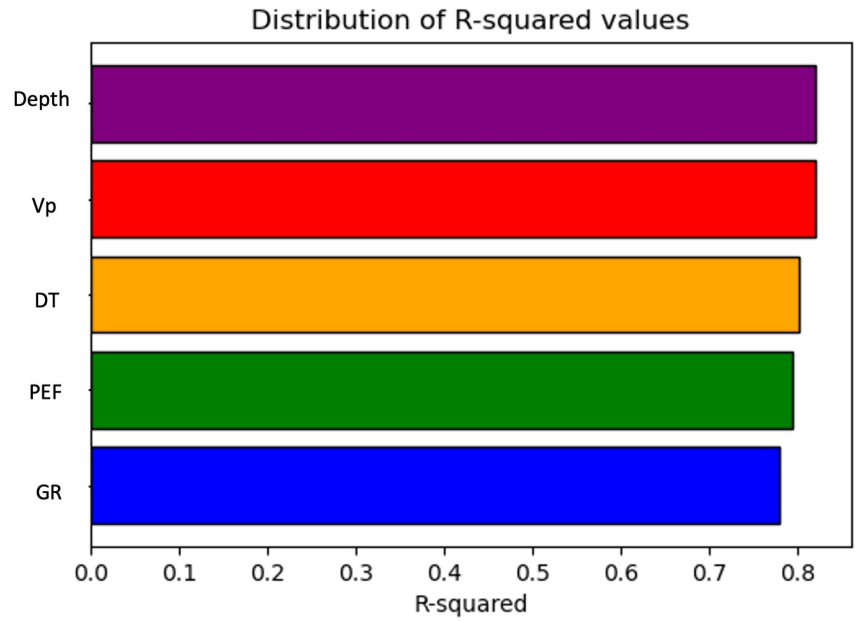

Figure 6. Results from the feature selection methods highlight the influential factors of selected logging curves for evaluating mineral volumes.

$$y = \sigma\left(\sum_{i=1}^{m} w_i x_i + b\right) \tag{13}$$

The activation function σ here is the rectified linear unit:

$$\sigma = x^+ = \max(0,x) = \frac{x + |x|}{2} = \begin{cases} 0 & \text{if } x > 0, \\ x & \text{otherwise.} \end{cases} \tag{14}$$

Equation 13 accommodates a single hidden layer and it was adapted to represent the MLP architecture employed in this study, characterized by a feedforward network encompassing a total of eight layers of nodes. This architecture includes one input layer with five input independent variables, six fully connected hidden layers, and ultimately, an output layer featuring four target dependent variables, as delineated in Figure 7. Each of the six hidden layers contains a specific number of neurons, namely 10, 20, 40, 20, 10, and 5, respectively. The set of equations that comprehensively characterizes the entire MLP architecture, as illustrated in Figure 7, and detailed below:

Table 3
R-Squared Values During Bidirectional Selection

Input variables	R^2
DEPTH	0.8195
Vp	0.8191
DT	0.8005
PEF	0.7933
GR	0.7789

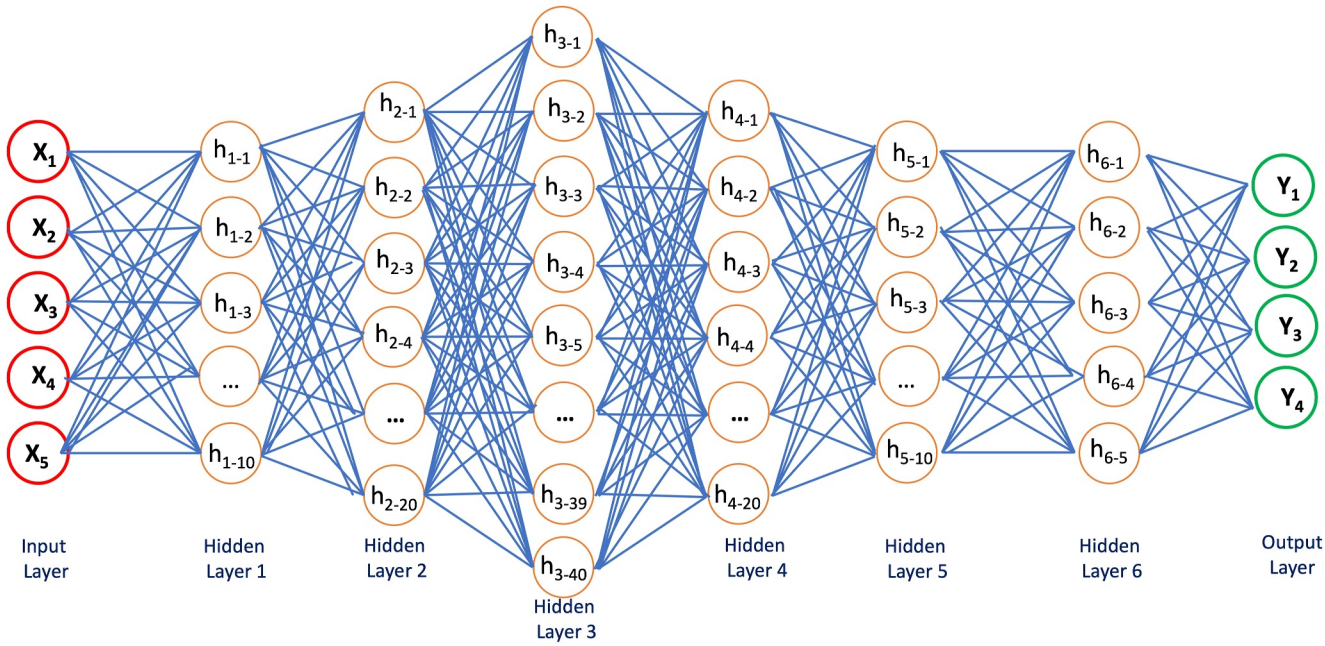

Figure 7. Multi-layer perceptron architecture with 8 layers of nodes. Comprising an input layer with 12 input variables, 6 hidden layers, and an output layer with 4 target variables.

$$
\begin{aligned}
h_i^{(1)} &= \sigma\left(\sum_{j=1}^{5} w_{ij}^{(1)} x_j + b_i^{(1)}\right), \quad i = 1,2,\ldots,10 \\
&\Downarrow \\
h_i^{(2)} &= \sigma\left(\sum_{j=1}^{10} w_{ij}^{(2)} h_j^{(1)} + b_i^{(2)}\right), \quad i = 1,2,\ldots,20 \\
&\Downarrow \\
h_i^{(3)} &= \sigma\left(\sum_{j=1}^{20} w_{ij}^{(3)} h_j^{(2)} + b_i^{(3)}\right), \quad i = 1,2,\ldots,40 \\
&\Downarrow \\
h_i^{(4)} &= \sigma\left(\sum_{j=1}^{40} w_{ij}^{(4)} h_j^{(3)} + b_i^{(4)}\right), \quad i = 1,2,\ldots,20 \\
&\Downarrow \\
h_i^{(5)} &= \sigma\left(\sum_{j=1}^{20} w_{ij}^{(5)} h_j^{(4)} + b_i^{(5)}\right), \quad i = 1,2,\ldots,10 \\
&\Downarrow \\
h_i^{(6)} &= \sigma\left(\sum_{j=1}^{10} w_{ij}^{(6)} h_j^{(5)} + b_i^{(6)}\right), \quad i = 1,2,\ldots,5 \\
&\Downarrow \\
y_i &= \sigma\left(\sum_{j=1}^{5} w_{ij}^{7} h_j^{(6)} + b_i^{(7)}\right), \quad i = 1,2,3,4
\end{aligned}
\tag{15}
$$

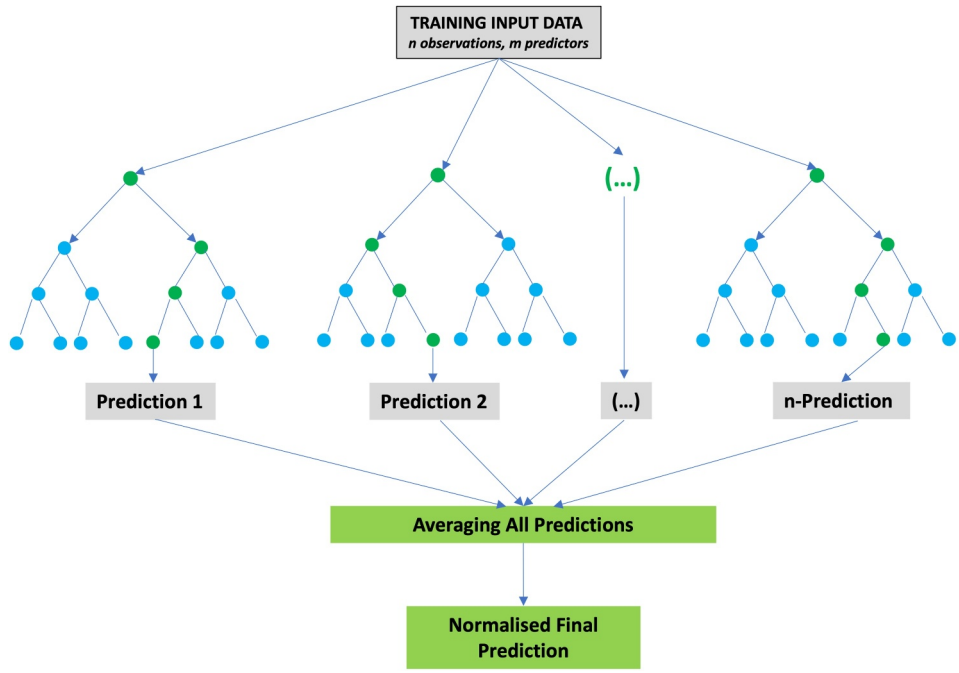

Figure 8. The architecture describing the random forest regression model used for this study.

where, $h_i^{(l)}$ represents the activation of the ith neuron in the hidden layer l (where l ranges from 1 to 7). $w_{ij}^{(l)}$ represents the weight connecting neuron i in layer l to neuron j in layer $l + 1$. The evaluation of the model prediction is based on the evaluation metrics defined in Section 3.2.4.

3.2.5.1. Normalization of Predictions

As captured in general ML workflow (Figure 3), the results from the ML algorithms are further normalized by the positivity and the unity constraints. Mathematically, the summation of minerals adheres to the general formula

$$V_{\text{norm}[i]} = \frac{V_i}{\sum_{i=1}^{n} V_i} \tag{16}$$

Here, $V_{\text{norm}[i]}$ signifies the normalized mineral volumes specific to the geological formations, V_i represents the volumes of individual minerals, and n denotes the number of minerals considered within the intervals. In this study, $n = 4$ for the Muschelkalk formations and $n = 2$ for the Buntsandstein formations. Drawing upon the primary mineral description outlined by Aichholzer et al. (2016, 2019), Equation 16 has been adapted to aid the determination of four minerals, namely calcite, dolomite, clay, and quartz, within the Muschelkalk formations. In the case of the Buntsandstein formations, this adapted equation focuses on the estimation of two minerals, clay and quartz.

3.2.6. Mineral Volume Inversion Using Random Forest Regression (RFR)

RFR is one of the most widely used ensemble supervised ML algorithms that combine multiple decision trees (Figure 8) for performing regression tasks with continuous target variables (Biau, 2012; Breiman, 2001; Pwavodi et al., 2023). Breiman (2001) proposed this ensemble method which independently builds each decision tree and trained on a random subset

$$\bar{r}(X, D_n) = E_\Theta[r_n(X, \Theta, D_n)] \tag{17}$$

where $X = (x_1, x_2, \ldots, x_n)$, Θ denotes a random subset of input features, D_n is the training data set and E_Θ denotes expectation with respect to the random parameter; it is introduced by selecting different subsets of features

(represented by Θ) and different data subsets (represented by $r_n(X, \Theta, D_n)$), r_n represents the prediction made by an individual decision tree within a random forest ensemble (Figure 8). In practice, the estimation of E_Θ is performed using Monte Carlo simulation (Biau, 2012). This involves generating a large number of random trees, typically denoted as M and computing the average of the individual outcomes (Biau, 2012). Each of the generated randomized trees $r_n(X, \Theta)$ outputs the average over all targets (y_i) (Figure 8) for which the corresponding vectors X_i fall in the same cell of the random partition as X:

$$r_n(X,\Theta) = \frac{\sum_{i=1}^{n} y_i 1_{[X_i \varepsilon A_n(X,\Theta)]}}{\sum_{i=1}^{n} 1_{[X_i \varepsilon A_n(X,\Theta)]}} 1_{\varepsilon_n(X,\Theta)} \tag{18}$$

When the expectation (E_Θ) is computed with respect to the random subset of input features (Θ) as shown in Equation 17, a modified version of Equation 18 is derived to estimate the RFR model

$$\bar{r}_n(X) = E_\Theta[r_n(X,\Theta)] = E_\Theta\left[\frac{\sum_{i=1}^{n} y_i 1_{[X_i \epsilon A_n(X,\Theta)]}}{\sum_{i=1}^{n} 1_{[X_i \epsilon A_n(X,\Theta)]}} 1_{\epsilon_n(X,\Theta)}\right] \tag{19}$$

where the event $\varepsilon_n(X, \Theta)$ in Equations 18 and 19 is defined using the following equation:

$$\varepsilon_n(X,\Theta) = \left[\sum_{i=1}^{n} 1_{[X_i \varepsilon A_n(X,\Theta)]} \neq 0\right] \tag{20}$$

Employing RFR for mineral volume estimation provides further insights into the interaction of variables. For every decision tree model used in this study, bootstrapping was carried out to randomly perform row and feature sampling. The choice of the best parameters for the RFR algorithm was done by a grid search tuning technique that attempts to compute the optimum values of the hyperparameters. The *max-depth* parameter for each of the decision tree is set to maximum depth of 20 nodes, the maximum leaf nodes and the number of decision trees (*n-estimator*) is set to 200. The evaluation of the model prediction is based on the evaluation metrics defined in Section 3.2.4. Subsequently, a parametric study was conducted to assess how the utilization of a greater depth of trees could impact both the performance and computational capacity of the model.

3.2.7. Mineral Volume Inversion Using Gradient Boosting Regression

Gradient boosting regression is an ensemble learning technique that combines the predictions of multiple weak learners, typically decision trees, to create a strong predictive model (Friedman, 2001, 2002). The governing equations for GBR are centered around the minimization of a cost function, often the MSE (Equation 9), with respect to the model's predictions (Friedman, 2001, 2002). The model starts with a single leaf and sums all the residuals

$$F_0(x) = \text{argmin}_\gamma\left(\sum_{i=1}^{n} L(y_i,\gamma)\right) \tag{21}$$

where $F_0(x)$ is the function of the input variable with (x) for the initial model, which we aim to improve iteratively, arg $\min_\gamma$ searches for the parameter γ that minimizes its argument, $L(y_i, \gamma)$ is the loss function applied to the target value y_i for the data points and the parameter γ. The loss function in Equation 21 quantifies how well the current model predicts the target value for each data point. The loss function with respect to the current prediction is calculated, representing the direction of steepest descent to minimize the loss.

$$L'\left(y_i,F(x_i)\right) = \frac{\partial L\left(y_i,F(x_i)\right)}{\partial F(x_i)}, \quad \text{for } i = 1,\ldots,n \tag{22}$$

where $L'(y_i, F(x_i))$ quantifies the difference between the true target values (y_i) and the current prediction $F(x_i)$. GBR is designed in such a way that it trains on the decision trees in a sequential manner with the next tree taking account of the residuals (r_i) of the previous decision tree.

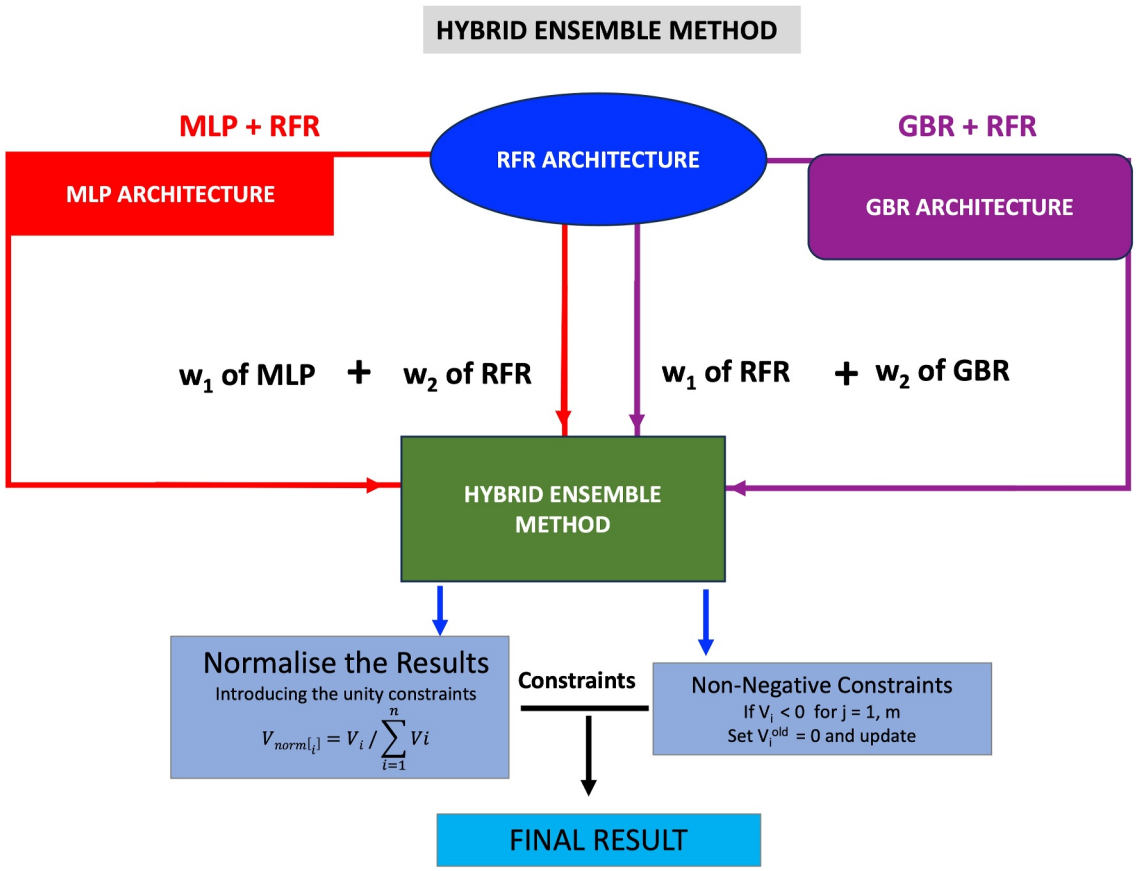

Figure 9. A hybrid ensemble model that combines proportions of multi-layer perceptron, gradient boosting regression, and random forest regression results.

$$r_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \quad \text{for } i = 1,\ldots,n \tag{23}$$

A smaller learning rate is introduced to control the contribution of each tree to the final prediction of the model $\left(F(x_i)_{new} = F(x_i)_{old} + \eta * T_i(x)\right)$. Finally to make the model's prediction more accurate, Equation 22 can be rewritten accounting for the residuals (Equation 23), hence taking account of previous predictions.

$$\gamma_m = \text{argmin}_\gamma \left(\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma)\right) \tag{24}$$

where γ_m represents the optimal value of γ at the mth boosting iteration. GBR was also used for mineral volume estimation because it is unbiased toward any specific feature type and can manage a high volume of features. The evaluation of the model prediction is based on the evaluation metrics defined in Section 3.2.4.

3.2.8. Mineral Volume Inversion Using Hybrid Ensemble Method

Although the predicted results (in terms of *MSE* and R^2-score) from the MLP, RFR, and GBR are encouraging, there is still room to improve the model's precision. Therefore, we propose a new weighted fusion process (Figure 9) that makes the use of these earlier results, with the condition that the combination of the weights (w_i) must be unity. Introducing w_i as the weight associated with the ith machine model, and y_i as the target result produced by that model

$$\sum_{i=1}^{n} w_i \cdot y_i \tag{25}$$

The selection of weights in Equation 25 can be achieved through a parametric study, employing an automatic random selection process within the range of 0–1. The quantity of samples (n) aligns with the number of target models (y_i) as defined in Equation 25. The normalization of weights ensures that their sum equals 1. To normalize, we simply substitute V_i in Equation 26 with w_i

$$w_{\text{norm}[i]} = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{26}$$

Here, we are dealing with just the results from the three ML algorithms (MLP, RFR and GBR) discussed earlier. In practice, we are combining Equations 13, 19 and 24,

$$y_{ens_1} = w_1 \cdot y_{MLP} + w_2 \cdot y_{RFR} \tag{27}$$

$$y_{ens_2} = w_1 \cdot y_{RFR} + w_2 \cdot y_{GBR} \tag{28}$$

$$y_{ens_3} = w_1 \cdot y_{MLP} + w_2 \cdot y_{GBR} \tag{29}$$

where y_{ens_x} is the final prediction from combining the three methods, y_{MLP}, y_{RFR}, and y_{GBR} are the predicted results obtained from Equations 13, 19, and 24 respectively. The determination of w_1 and $w_2 = 1 - w_1$ values were based on the parametric study as discussed above, employing a random selection process to iteratively find the optimal weighting parameters. This process involved generating up to 100 unique values, each time the simulation was run. The algorithm is written in a way to automatically select the optimal n-value for each iteration, based on the model's performance metrics such as the highest R^2 score and lowest MSE values. Subsequently, the models were utilized to predict the blind wells, and Equation 26 was applied to normalize the final mineral volumes.

4. Results

The methodologies employed in this study have been designed to precisely characterize the distribution of mineral volumes within the Muschelkalk and Buntsandstein formations. We present here the outcomes of both statistical approaches employed for wells GRT-1 and OBR-101, and the ML methods applied to wells GPK-1 (Muschelkalk formation) and EPS-1 (Buntsandstein formation). The quantitative results derived from the ML models, assessed through various evaluation metrics, are also systematically presented.

4.1. Muschelkalk Formation

4.1.1. Mineral Volume Distribution Using the Statistical Approach

The application of the statistical method for estimating mineral volumes was first applied to the OBR-101 and GRT-1 wells. To ensure the method's accuracy and relevance, we took into account the mineral distribution evidence derived from core samples, cuttings, and the gamma-ray (GR) log descriptions as provided in the work of Aichholzer et al. (2016, 2019). Figure 10 provides a visual representation of the mineral content within the Muschelkalk formation of the OBR-101 and GRT-1 wells. Five distinct minerals were identified in this formation: calcite, clay, dolomite, anhydrite, and quartz. In the upper and middle Muschelkalk layers (Figure 10), four dominant minerals are observed: calcite, clay, dolomite, and anhydrite. In contrast, in the lower Muschelkalk, three primary minerals are dominant: dolomite, clay, and quartz.

For ease of reference, based on mineral volume distribution, we have reclassified the Muschelkalk into two main intervals: the top Muschelkalk interval, encompassing the Upper and Middle Muschelkalk and the Lower Muschelkalk. Proportional analysis indicates that calcite is the predominant mineral within the top Muschelkalk, followed by the clay. A prominent feature in both wells is the presence of an anhydrite-rich zone, recognized as the "Marnes Barriolées" (Aichholzer et al., 2016, 2019). Within the Lower Muschelkalk, the mineral proportions are generally characterized by a dominance of dolomite, clay, and quartz, respectively. The outcomes of the

10.1029/2024JH000154

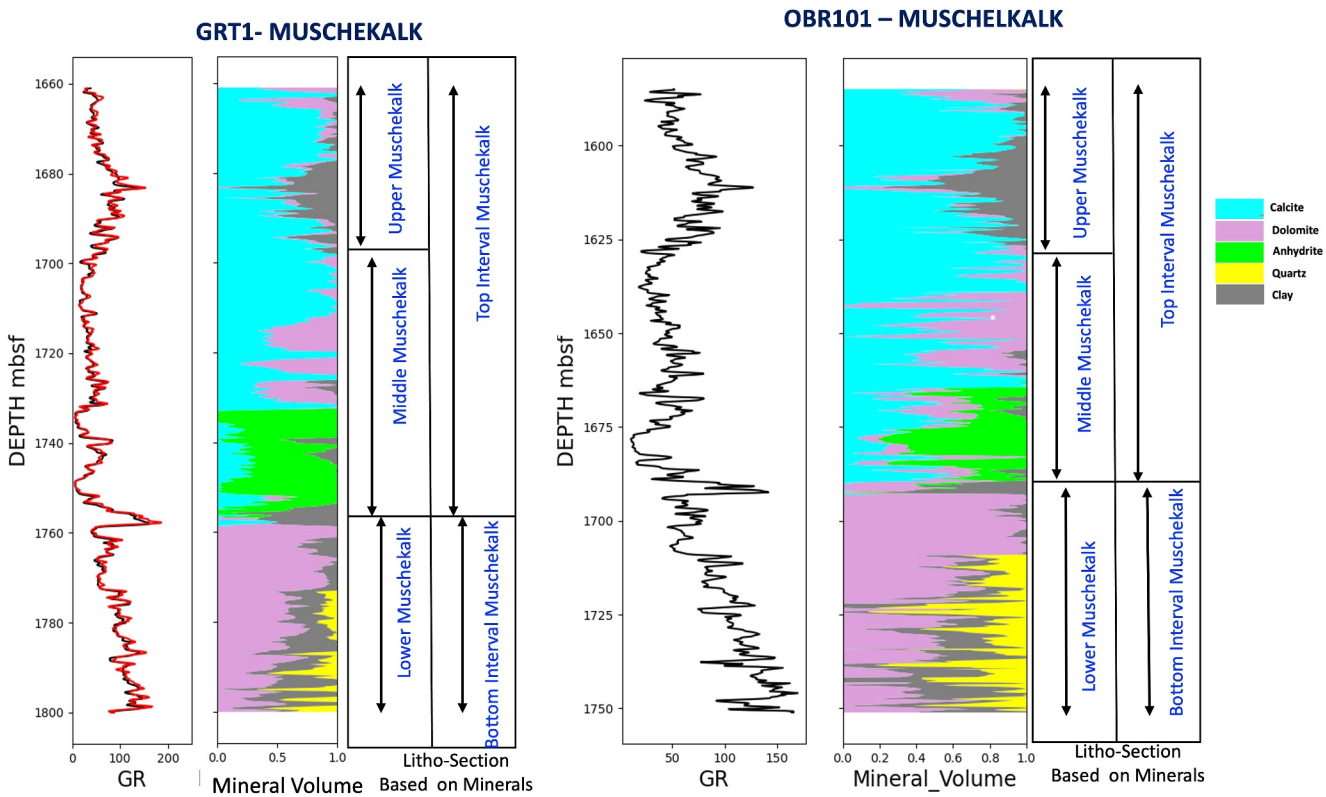

Figure 10. Mineral volume estimates of calcite, dolomite, anhydrite, clay, and quartz within the Muschelkalk in GRT-1 and OBR-101 wells, using the statistical method.

statistical method have yielded a credible estimate of mineral proportions within the Muschelkalk formations at the OBR-101 and GRT-1 wells.

4.1.2. Mineral Volume Distribution Using Machine Learning

We attempt to predict the mineral volume distribution within the Muschelkalk formations along the GPK-1 well with ML models, using the results above from the OBR-101 and GRT-1 wells serving as dependent variables in the training process. These models reveal the presence of four dominant minerals in the Muschelkalk formations around GPK-1: calcite, dolomite, clays, and quartz, as depicted in Figure 11. We observe a slight variation in the mineral compositions between GPK-1 and OBR-101/GRT-1 wells. The distinctive interval characterized by anhydrite (Marnes Bariolées), observed in the OBR-101 and GRT-1 wells, is not observed in the GPK-1 well, aligning with the detailed lithological classifications conducted by Aichholzer et al. (2016, 2019) for GPK-1.

Figure 11 shows that the Top Muschelkalk is predominantly composed of calcite, dolomite, and clays. In contrast, the Lower Muschelkalk is dominated by dolomite and clay, followed by quartz, with a minimal fraction of calcite at the top part of this interval. The assessment of the ML models' accuracy and performance is further elaborated below, focusing on the four key evaluation metrics, as described in Section 3.2.4.

For the calcite mineral volume (Table 4), the RFR model exhibits superior performance, with the lowest MSE, the highest R^2 score and the lowest MAE values. Following closely is the GBR + RFR hybrid ensemble model. Both RFR and GBR + RFR models demonstrate minimal bias, reflected in the lowest sum and mean of residuals, indicating consistent predictions closely aligned with actual values. Overall, the evaluation metrics (Table 4) document the superior performance of the RFR and GBR + RFR models, surpassing MLP, GBR, and MLP + RFR models. Overall, the cross-plots between the actual and predicted calcite volumes for all the methods are close to a linear trend (Figure 12). However, for values smaller than 0.6, the predicted values from the MLP and GBR algorithms tend to be overestimates. This observation suggests that while the models generally perform well and align closely with the actual values, they may overpredict the calcite volume, particularly at lower fractions.

10.1029/2024JH000154

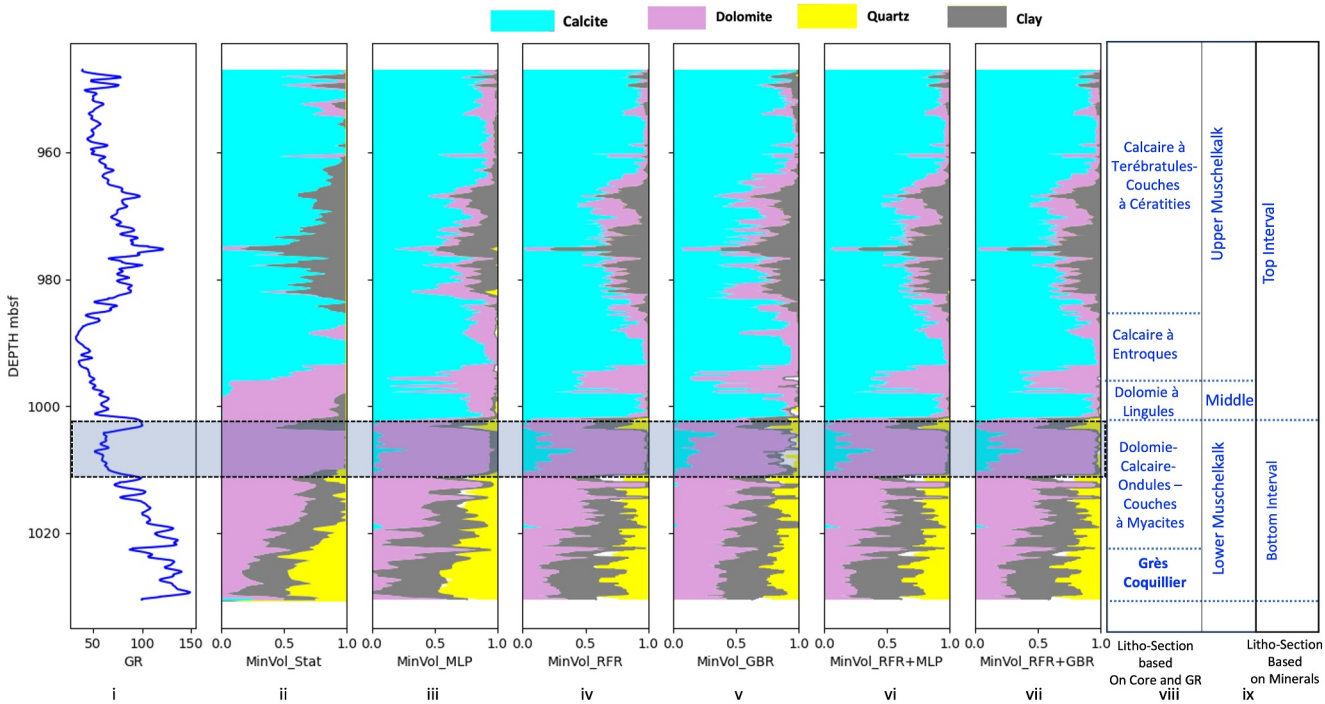

Figure 11. The distribution of estimated minerals volume using the machine learning algorithms within the Muschelkalk interval along well GPK-1.

For clay volumes (Table 4). Here again, the GBR + RFR and RFR models show the best performance with the lowest MSE, the highest R^2-scores, the lowest MAE values and the highest EVS values. These two models consistently outperform the other models across all evaluated metrics, with exceptionally low prediction errors, high predictive power, and outstanding accuracy in capturing the variance in the data. While other models, such as MLP, also perform well, they exhibit slightly higher errors and slightly less predictive power compared to RFR and GBR + RFR. Figure 12 shows a strong linear trend between the predicted and actual clay volumes for all the methods.

For the dolomite, the results presented in Table 4 and Figure 12 demonstrate that the RFR and GBR + RFR models exhibit superior performance. The cross-plots of the observed versus predicted dolomite volumes show a linear trend between them but again with some overestimates at small volume fractions. The conclusions for quartz are the same as for dolomite (Figure 12): here again (Table 4), the RFR and GBR + RFR models offer higher accuracy in their predictions and exhibit fewer errors when estimating the proportion of quartz. However in Figure 12, we point out that all methods systematically underestimate the quartz volume when it is greater than 0.4.

4.2. Buntsandstein

We apply the same workflow for the much simpler case of the Buntsandstein. Indeed, as stated above, there are only two minerals of interest: quartz and clay (Figures 13 and 14). We use results from well GPK-1 as training sets for predicting mineral volumes within the Buntsandstein along EPS-1. The metrics from the different ML models used for the Buntsandstein formations are presented in Table 5. Similarly to what we observed in the Muschelkalk, the RFR, GBR + RFR and MLP + RFR ensemble models consistently perform better, with the lowest MSR and highest R^2-scores. The performance of the models in terms of the estimated mineral volumes is further compared (Figure 13) with the ground truth information based on the XRD analysis of 15 samples along the Buntsandstein formations (Heap et al., 2017, 2019).

Table 4
The Evaluation Metrics of the Calcite, Clay, Dolomite, and Quartz Minerals for All the Machine Learning Methods

Minerals	Metrics	MLP	RFR	GBR	MLP + RFR	GBR + RFR
Calcite	MSE	0.04192	0.0124	0.0216	0.0027	0.0209
	R^2-score	0.7096	0.8521	0.8503	0.8426	0.8554
	MAE	0.1512	0.10019	0.0976	0.1048	0.0966
	EVS	0.7102	0.8521	0.8506	0.8428	0.8556
	Sum of residual	3.5420	0.5603	2.4127	0.9958	0.8332
	Mean of residual	0.0088	0.0014	0.0060	0.0025	0.0021
Clay	MSE	0.0038	0.00269	0.0023	0.0027	0.0023
	r^2-score	0.887618	0.9214	0.9313	0.9225	0.9327
	MAE	0.0419	0.0337	0.0318	0.0327	0.0315
	EVS	0.8880	0.9213	0.9315	0.9225	0.9328
	Sum of residual	1.4597	0.3632	0.7129	0.1015	0.3895
	Mean of residual	0.0036	0.0009	0.0018	0.0003	0.0010
Dolomite	MSE	0.0282	0.0189	0.0190	0.0184	0.0180
	r^2-score	0.7029	0.8197	0.8008	0.8058	0.8105
	MAE	0.1232	0.0995	0.0976	0.1000	0.0962
	EVS	0.7040	0.8101	0.80090	0.8062	0.8105
	Sum of residual	4.1415	2.0428	0.6858	2.2672	1.8241
	Mean of residual	0.0103	0.0051	0.0017	0.0057	0.0045
Quartz	MSE	0.0070	0.0046	0.0045	0.0047	0.0044
	r^2-score	0.5670	0.7180	0.72052	0.7095	0.7278
	MAE	0.0380	0.0244	0.0270	0.0256	0.0260
	EVS	0.5691	0.7184	0.7211	0.7104	0.7285
	Sum of residual	2.3540	0.9584	1.3417	1.0920	0.9139
	Mean of residual	0.0058	0.0024	0.0033	0.0027	0.0023

4.3. Model Performance Based on Parametric Study

4.3.1. RFR Architecture

In Section 3.2.6, the rationale behind choosing the RFR architecture was based on the outcomes of a grid-search tuning process that involved parameters such as *max-depth*, *leaf nodes*, and *n-estimators*. Figure 15 sheds light on the impact of varying *max-depth* values from 1 to 400, with a constant number of decision trees (*n-estimators*) set at 200. The results indicate that the random forest ensemble emerges as a powerful and expressive model, adept at capturing diverse patterns and generalizations within the data. Significantly, the analysis reveals that, for all minerals, beyond a *max-depth* of 20 (Figure 15), both the training and validation models tend to exhibit a slight tendency to overfit or underfit. The training model stabilizes notably after reaching a *max-depth* of 100 (Figure 15), while the behavior of the validation model varies for different minerals (Figure 15).

Specifically, the validation model for clays (Figure 15) remains stable between 100 and 300 before under-predicting, for calcite (Figure 15) stabilizes and then overfits (from a max-depth of 300), while the validation model for quartz displays over- or under-predictions (Figure 15). In contrast, the validation model for Dolomite appears relatively stable (Figure 15). Considering the overall observations derived from the evaluation metrics, the RFR method consistently proves to be effective, demonstrating a robust performance across all four minerals. However, careful consideration is needed to address potential challenges, such as the risk of overfitting associated with excessive use of deeper trees and increased computational complexity during training and evaluation. Achieving a balance through empirical experimentation and validation on separate data sets is imperative for optimizing the model's performance and ensuring robust generalization to the test data.

10.1029/2024JH000154

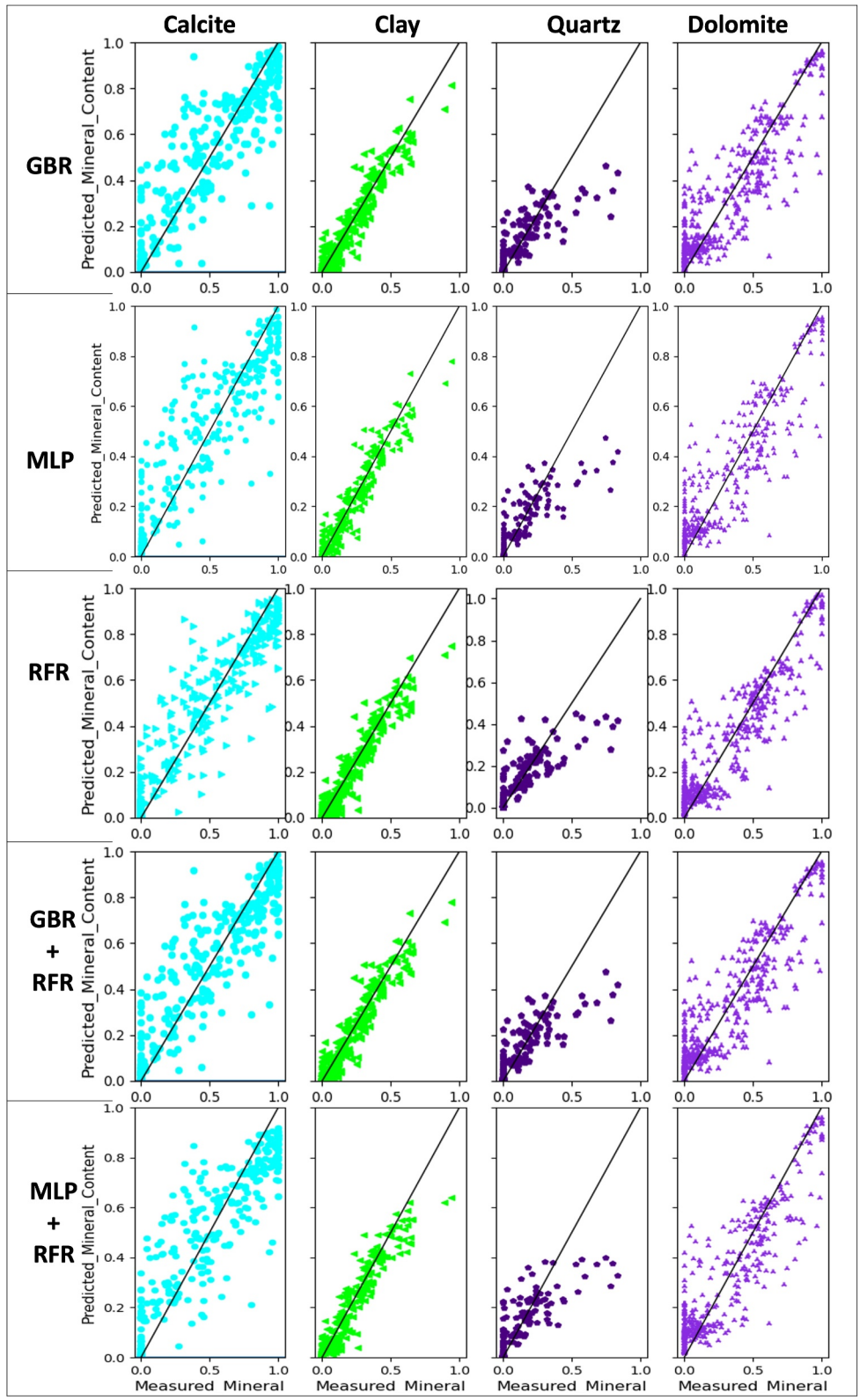

Figure 12. Cross-plots of measured and predicted values of the four identified minerals within the Muschelkalk formations (calcite mineral, clay, quartz, and dolomite mineral volumes) for each machine learning methodology.

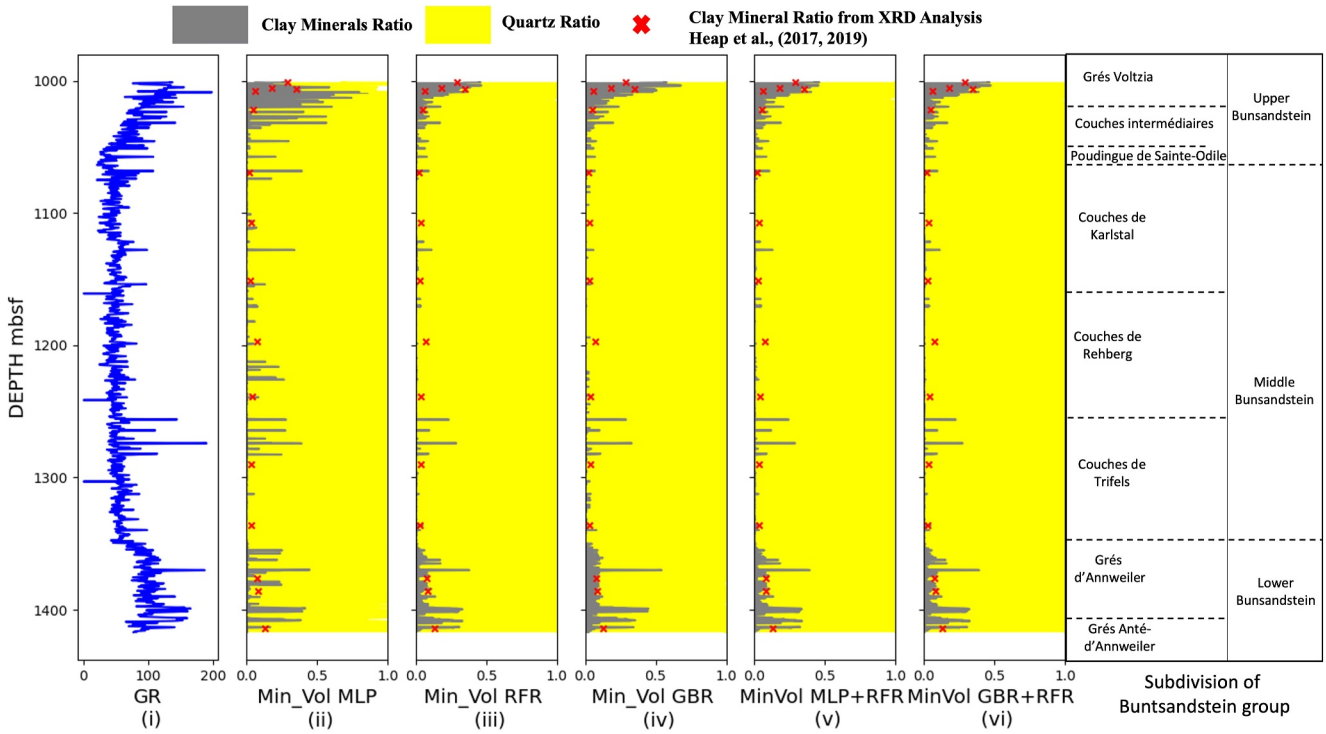

Figure 13. The distribution of estimated mineral volume using the machine learning algorithms within the Buntsandstein formations along well EPS-1. The red x symbol represents the ratio of the clay minerals from X-ray diffraction analysis from Heap et al. (2017), Heap et al. (2019).

4.3.2. Hybrid Ensemble Architecture

The proposed hybrid ensemble model (Figure 9) introduces a weighted fusion process that combines the predictions of three machine-learning algorithms (MLP, RFR, and GBR). One advantage of this hybrid ensemble approach lies in its ability to allow for a transition between different weight configurations, facilitating adaptability to the specific strengths and weaknesses of each algorithm and resulting in a more robust prediction. We performed a parametric study by randomly varying the weights (Equations 27–29).

The findings for the MLP + RFR ensemble (Figure 16) demonstrate a significant pattern, especially in relation to clay (Figure 16a), where the maximum R^2 scores are achieved with ideal *n-values* ranging from 0.82 to 0.86 (Figure 16a). We note (Figure 16a) that it is possible to improve the performance of the clay model by choosing *n-values* between 0.82 and 0.86 for RFR, while assigning the remaining to MLP. There is a strong bias toward RFR when it comes to calcite, quartz, and dolomite (Figures 16b–16d), when the *n-value* is close to 1. There is, however, a general decline in model performance as the emphasis shifts toward MLP. Increasing the *n-values* of MLP does not yield significant improvement and tends to decrease the overall model performance. This observation emphasizes the interplay between model components and underscores the importance of a sound approach in achieving optimal performance. In general for the MLP + RFR ensemble with the right *n-value* range, the model performance can be improved.

For the GBR + RFR ensemble approach (Figure 17) the model performance for the four minerals exhibits a parabolic shape. The vertex of the parabolic shape of the scatter model points (Figure 17) represents the most optimal model (Figure 17), while along the arms of the parabolas, whether toward RFR or GBR, lower performing models are observed (Figure 17). In Figure 17a, the best performing model for clay is biased toward GBR within the *n-values* range of 0.24–0.30. Beyond this range, higher *n-values* introduce a bias toward RFR (Figure 17a), but with a decline in overall model performance. On Figures 17a–17c, the model performance for calcite, quartz, and dolomite minerals seems to be quite similar. The model with the best performance for calcite (Figure 17b) has *n-values* ranging from 0.37 to 0.53. The optimal range of *n-values* for quartz is 0.45–0.55 (Figure 17c). In contrast, for dolomite the best performing models are associated with *n-values* in the range of 0.55–0.72 (Figure 17d).

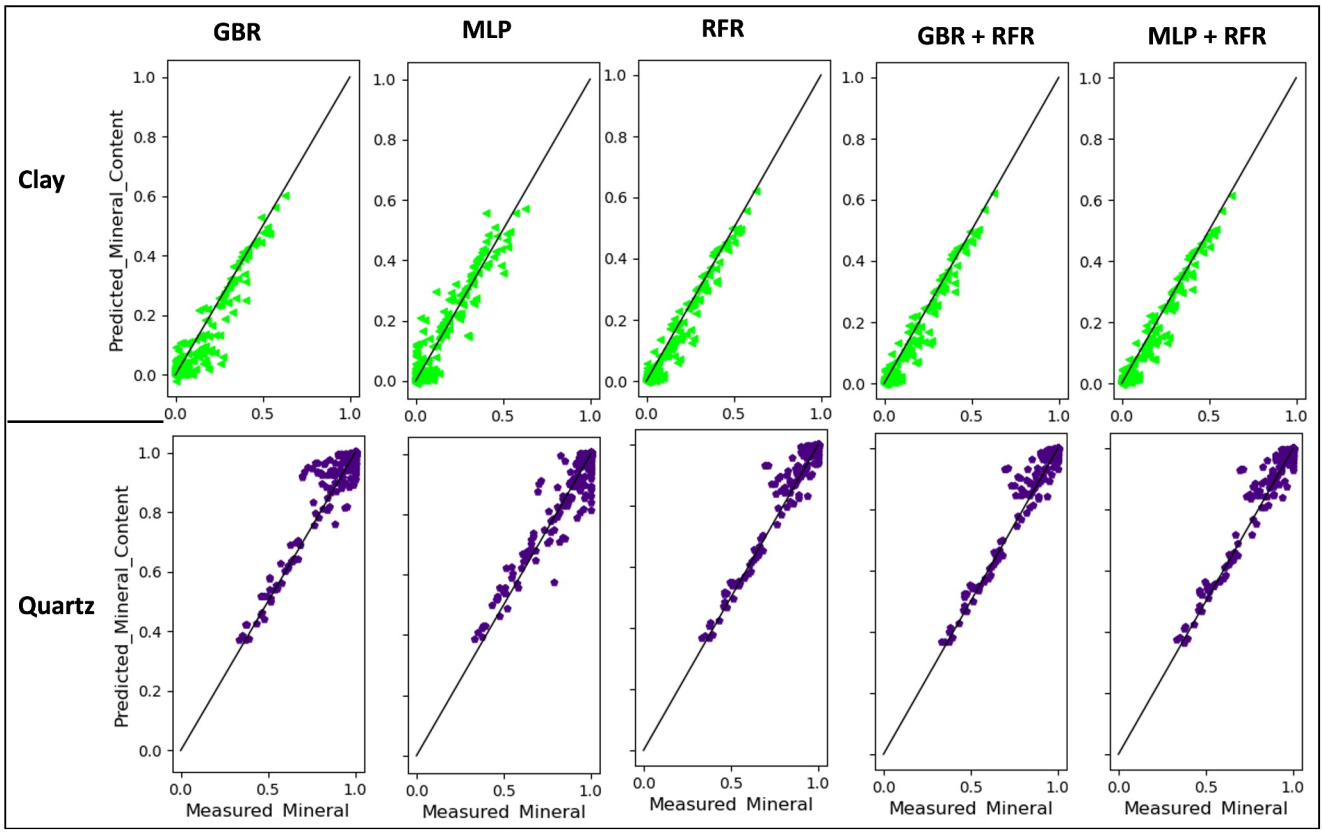

Figure 14. Cross-plots of measured and predicted values of the two identified minerals within the Buntsandstein formations (quartz and clay volumes) for each machine learning methodology.

Qualitatively assessing the performance of the four minerals based on the spread of the parabolic height (Figure 17), it appears that quartz (Figure 17c) stands out as the most enhanced model.

In Figures 16 and 17, we show that the optimal *n-values* vary for each mineral within the multivariate regression model. For example, the *n-value* that yields the best model for clay differs from that of other minerals. Consequently, the multivariate system that relies on a single *n-value* for all minerals puts some of them at a

Table 5
The Evaluation Metrics of Quartz and Clay for All the Machine Learning Methods Used Within the Buntsandstein

Minerals	Metrics	MLP	RFR	GBR	GBR + RFR	MLP + RFR
Clay	MSE	0.0003	0.00011	0.00028	0.00011	0.00011
	r^2-score	0.9333	0.9743	0.9350	0.9747	0.9744
	MAE	0.0060	0.0032	0.0052	0.0032	0.0033
	EVS	0.9344	0.9743	0.9354	0.9747	0.9744
	Sum of residual	4.2919	0.2942	0.2739	0.342	0.2077
	Mean of residual	0.0021	0.0001	0.0001	0.0002	0.0001
Quartz	MSE	0.0004	0.0002	0.0004	0.0002	0.0002
	r^2-score	0.9278	0.9551	0.9163	0.9553	0.9560
	MAE	0.0062	0.0040	0.0062	0.0040	0.0041
	EVS	0.9278	0.9551	0.9163	0.9553	0.9560
	Sum of residual	0.7455	0.8522	0.3201	0.8044	0.4822
	Mean of residual	0.0004	0.0004	0.0002	0.0004	0.0002

10.1029/2024JH000154

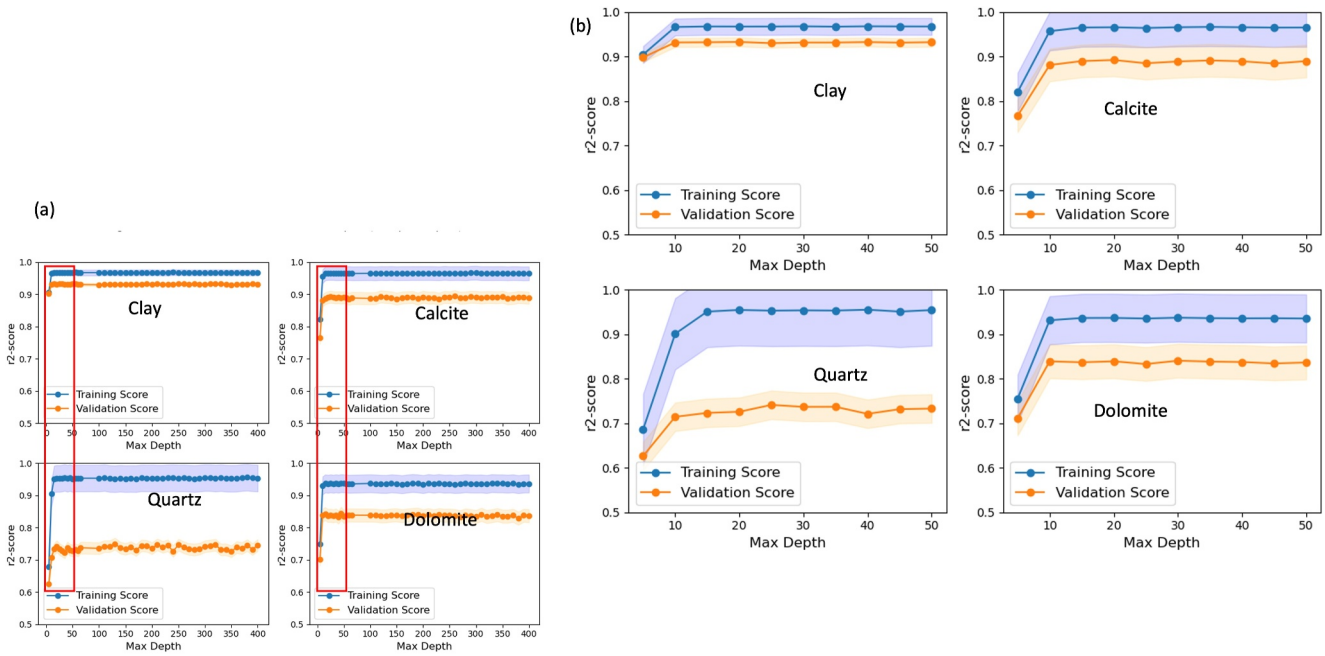

Figure 15. (a) Outcome of Hyperparameter Tuning Experiment for a random forest regression Algorithm, Examining various max-depth values and evaluating r^2-score across different minerals (clay, calcite, quartz, and dolomite) within the Muschelkalk Formation along Well GPK-1. (b) The top-right subplot provides a zoomed-in view of a specific region of interest indicated by the red rectangle in the main figure.

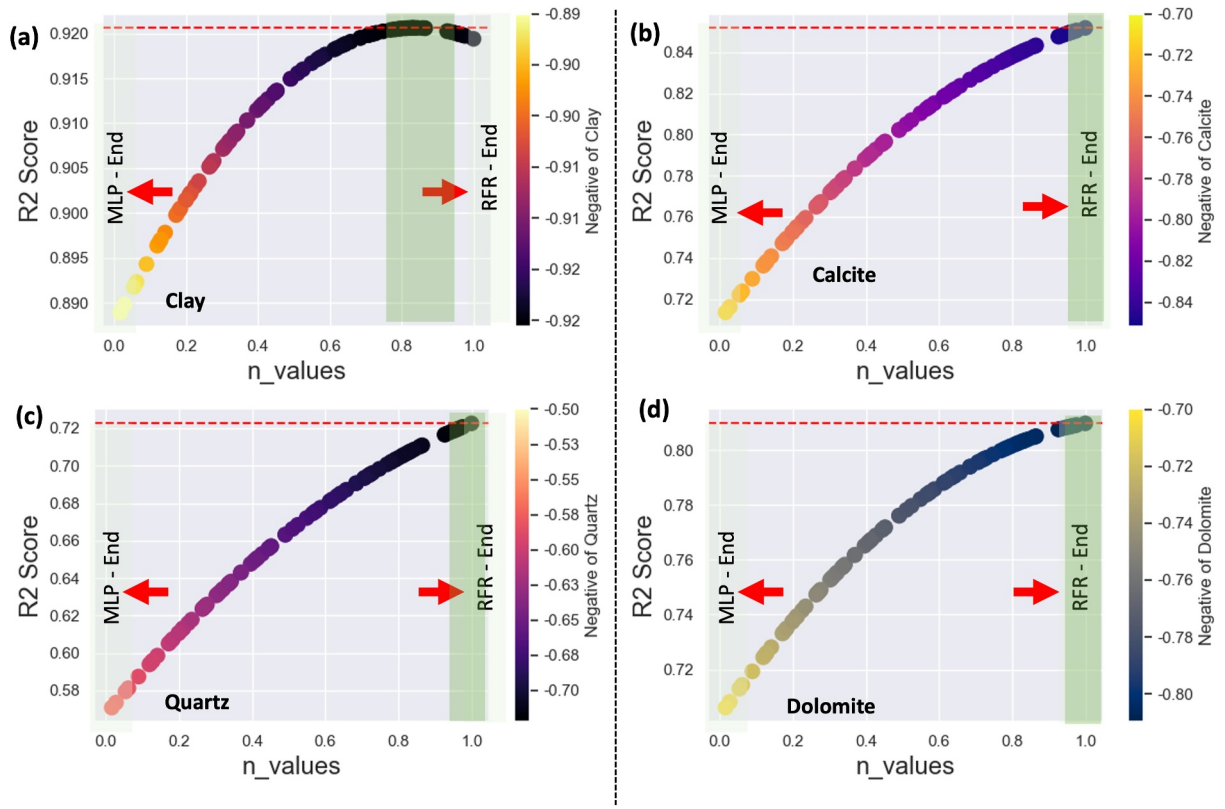

Figure 16. Outcome of parametric study after randomly generating 100 n-values for obtaining the most optimal multi-layer perceptron + random forest regression hybrid ensemble for all four minerals (clay, calcite, quartz, and dolomite) within the Muschekalk formations along Well GPK-1.

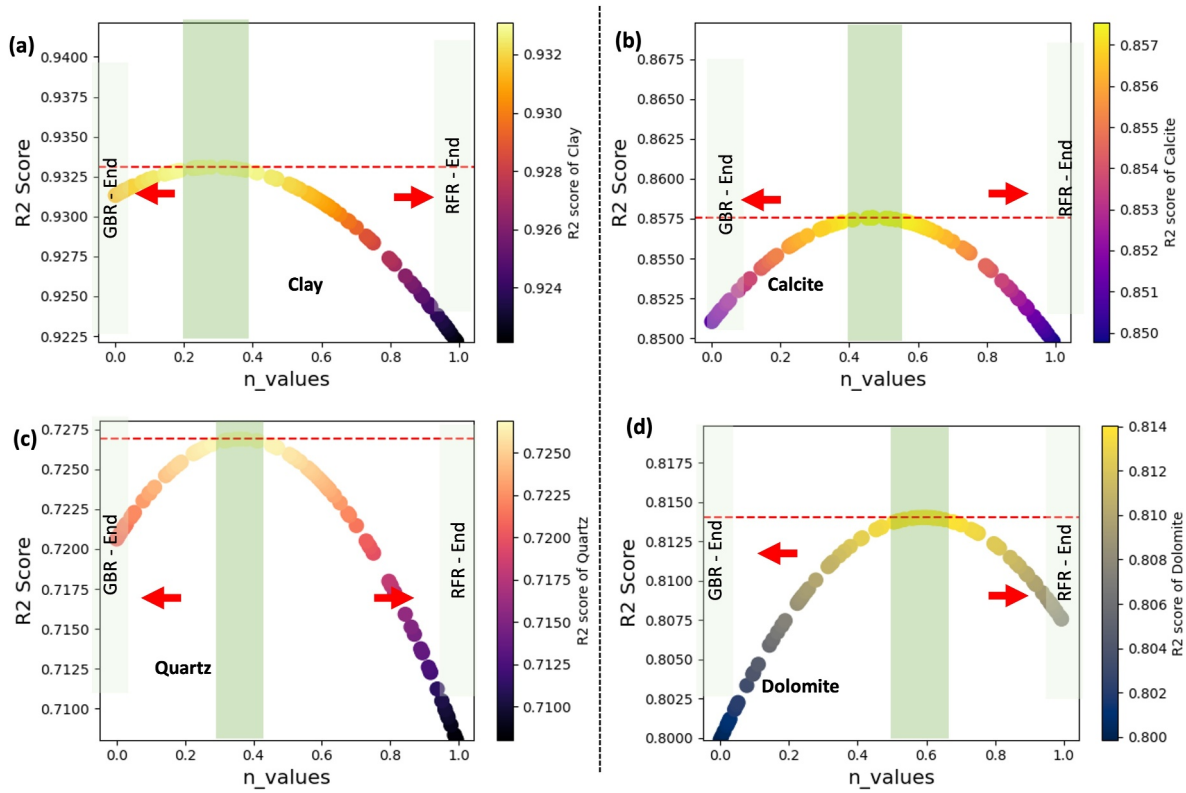

Figure 17. Outcome of parametric study after randomly generating 100 *n*-values as weights for obtaining the most optimal gradient boosting regression + random forest regression hybrid ensemble for all four minerals (clay, calcite, quartz, and dolomite) within the Muschekalk formations along Well GPK-1.

disadvantage. To address this issue, we conducted an additional analysis employing a univariate regression system. This involved considering unique *n-values* that resulted in the best optimal model for each mineral, hence, Equations 15, 19 and 24 were adapted for univariate regression, solving for different minerals individually. The GBR + RFR model was tested within this univariate regressional system, and quantitatively, the results indicate improved metrics in terms of R^2-score and MSE (Table 6 and Figure 18) for each mineral within the formations of interest. This underscores the advantage of adopting a mineral-specific approach in optimizing univariate regression models as opposed to relying on a generalized multivariate regression system. Overall, the parametric studies (Figures 16–18) highlight the relationship between *n-values*, model performance, and the use of a univariate regression system to solve individual minerals in the formations of interest. These implications emphasize the trade-off involved in making these choices. We will like to note that the above validation primarily focused on the Muschelkalk formations due to their higher mineral diversity compared to the simpler Buntsandstein

Table 6
Comparison of the Mean Squared Error and R^2-Score Performance of the Multi-Regression (MR) and Univariate Regression (UR) Algorithms

Minerals	Metrics	GBR + RFR − MR	GBR + RFR − UR	Diff = UR − MR
Calcite	MSE	0.02087	0.01811	−0.00276
	R^2-score	0.85541	0.87455	0.01914
Clay	MSE	0.00230	0.00226	−0.0004
	R^2-score	0.93271	0.93384	0.00113
Dolomite	MSE	0.01799	0.01628	−0.00171
	R^2-score	0.81046	0.82837	0.01791
Quartz	MSE	0.00442	0.00439	−0.00005
	R^2-score	0.72779	0.72954	0.00175

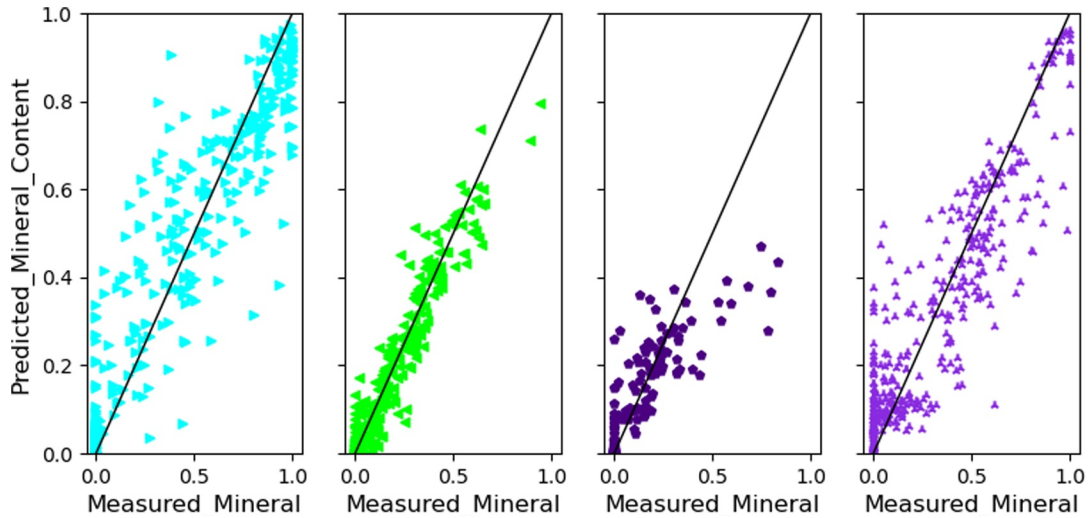

Figure 18. Results of the gradient boosting regression + random forest regression univariate regression: cross-plot of measured and predicted values of the minerals.

sandstone formations (two minerals), it's imperative to note that the hybrid method's technical application remains applicable to both scenarios.

4.3.3. Sensitivity Analysis of the Input Parameters on the Machine Models Output

The sensitivity study of input parameters to the overall model performance of MLP, RFR, and GBR machine algorithms is critical in determining their predictive dynamics. The first-order Sobol index, which elucidates the fraction of total output variation attributed to particular input parameters while keeping others constant, reveals their relative importance. Observations reveal notable influences: for the MLP machine algorithm (Figure 19a), the VP, DT, and GR have an impact on the clay volume ratio, while only the GR log retains a pronounced effect on the clay and quartz volume across RFR and GBR models (Figures 19b and 19c). In contrast, Figures 19b and 19c show that the Vp and DT have a minor influence on various minerals in GBR and MLP ML models.

5. Discussion

5.1. Mineral Volume Inversion: Statistical Approach Versus Machine Learning

In scrutinizing the mineralogical composition of the Muschelkalk formations, we observe a distinction between its upper and lower intervals, as unveiled by both statistical methods and advanced ML algorithms (Figure 11). Within the upper Muschelkalk, there is a pronounced prevalence of calcite, dolomite, and clay. In contrast, the lower Muschelkalk intervals contain dolomite, clay, and quartz. Interestingly, the ML algorithms detected a minor amount of calcite in the upper portion of the lower Muschelkalk (Figure 11), a finding that was not observed in the statistical analysis. This discrepancy may be attributed to the inherent differences in the methodologies employed. The statistical approach relies on a priori geological knowledge, guided by the primary rocks of the interval, alternating marls, marly dolomites, dolomites, shelly sandstone, and dolomitic sandstone interbedded with clayey marl beds (Aichholzer et al., 2016, 2019; Duringer et al., 2019).

Conversely, the ML algorithms operate independently of such predefined geological constraints. Instead, their functionality is governed by an ability to discern nonlinear relationships within diverse input data sets. The variation in technique discussed above provides insight into the differences seen, highlighting how both methodologies complement each other in understanding the complex mineralogical patterns within the Muschelkalk formations. The accuracy of each method can only be verified by a comparison to XRD estimated mineralogical composition of core samples. Unfortunately, there are no XRD analyses results for well GPK1's Muschelkalk formations.

On the other hand, a thorough analysis was conducted on the eight subdivisions of the Buntsandstein formation in well EPS1, as defined by Aichholzer et al. (2019), with particular emphasis on the evaluation of clay content and

10.1029/2024JH000154

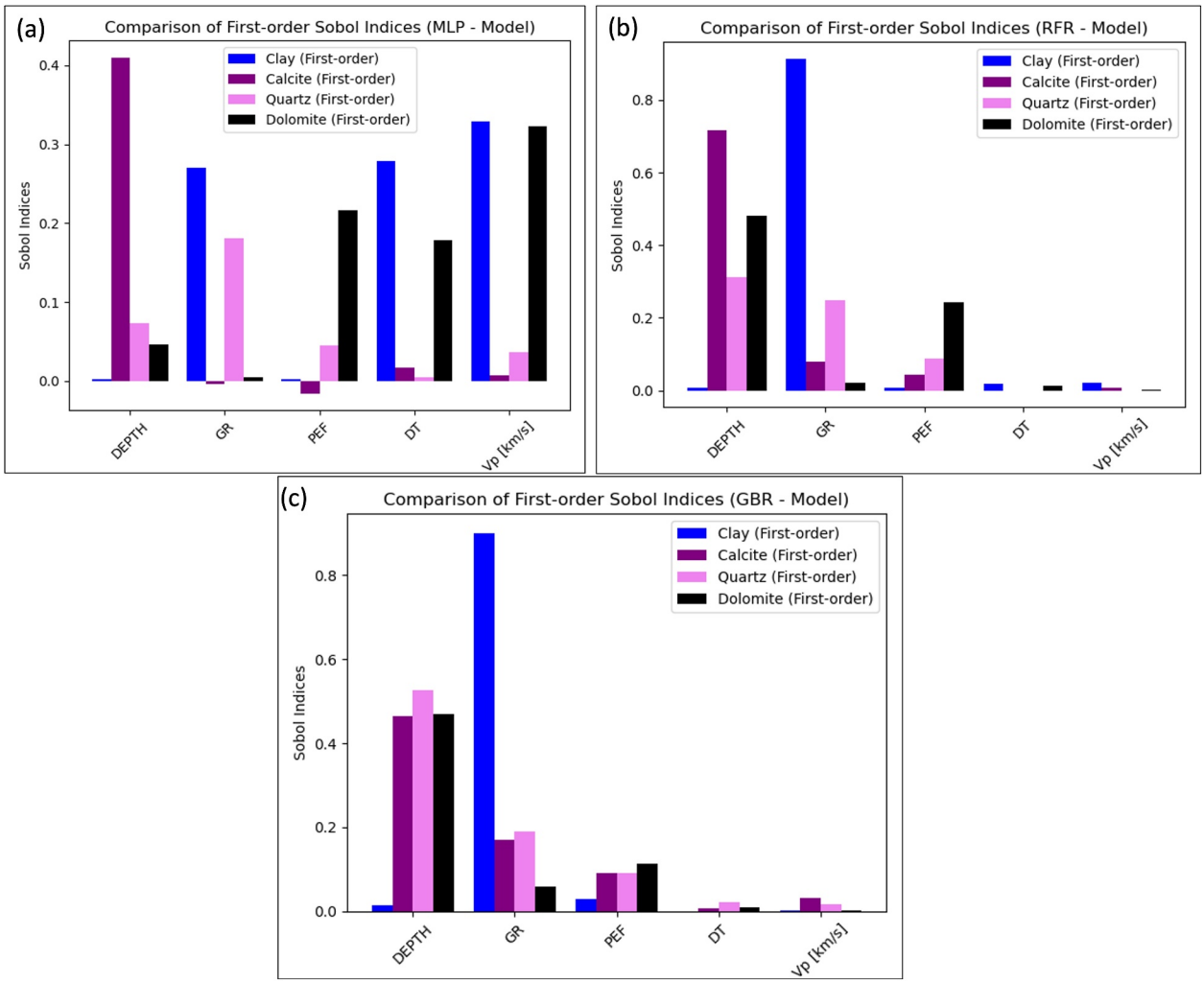

Figure 19. This histogram illustrates the comparison of first-order Sobol indices for independent input variables. Each vertical bar represents the magnitude of the Sobol index, indicating the relative importance of the corresponding item in explaining the variance of the model output. Four distinct colors represent the contribution of different minerals: blue for clay, purple for calcite, violet for quartz, and black for dolomite.

quartz properties. The objective of this first-order assessment was to determine which of the five ML algorithms (Table 5) effectively represents the mineralogical description in accordance with XRD analysis (Heap et al., 2017, 2019) and geological description (Aichholzer et al., 2019). Examining the findings illustrated in Figure 13 along the stratigraphic sequence of Buntsandstein formations, specifically from the upper to the lower Buntsandstein intervals, reveals consistent patterns across the ML methodologies. Notably, thick sandstone intervals, as depicted in Figure 13, are systematically interspersed with discontinuous centimetric to decametric clay layers. This alignment notably corresponds to the detailed descriptions provided in cores and cuttings by Aichholzer et al. (2019). Furthermore, in attempts to validate the accuracy of the ML models, the results obtained from XRD analyses (Table 7) sourced from studies conducted by Heap et al. (2017, 2019) were meticulously compared with the volume estimations of clay and quartz (red crosses in Figure 13).

A first-order analysis indicates an agreement between the volumetric assessments of clay (Figure 13) derived from various ML methods and the XRD data from the 15 core samples (Table 7). Upon delving deeper into the methodological outcomes, a nuanced observation emerges regarding their mutual complementarity. For instance, within the upper Buntsandstein interval (comprising Grès Voltzia, Couches intermédiaires, and Poudingue de Sainte-Odile) as well as the lower Buntsandstein interval (Grès d'Annweiler and Grès Ante d'Annweiler), the RFR, MLP + RFR, and GBR + RFR models (Figures 13iii, 13v, and 13vi) demonstrate a reasonable fit with the

Table 7
Quantitative Bulk Mineralogical Composition From X-Ray Powder Diffraction (X-Ray Diffraction) Analysis of 15 Sandstone Samples (Buntsandstein Formations) From Exploration Well EPS-1 at the Soultz-Sous-Forêts Geothermal Site (Alsace, France) Modified From Heap et al. (2019) (First Three Values) and Heap et al. (2017) (Others)

TVD (m)	Quartz (wt.%)	Clay (wt.%)
1,001	58	29
1,005.5	66	18
1,006.5	46	35
1,008	74.5 ± 1.6	6.0 ± 2.9
1,022	78.9 ± 1.7	5.0 ± 2.5
1,069	89.2 ± 0.4	2.0 ± 0.8
1,107	89.0 ± 1.1	3.2 ± 1.3
1,151	90.7 ± 1.2	2.8 ± 1.3
1,197	83.4 ± 2.6	7.3 ± 3.2
1,239	87.8 ± 1.3	3.8 ± 1.5
1,290	86.7 ± 1.6	3.5 ± 2.1
1,336	82.3 ± 1.7	3.0 ± 1.8
1,376	73.3 ± 3.0	7.8 ± 3.9
1,386	70.6 ± 2.8	8.3 ± 4.5
1,414	66.4 ± 4.0	13.1 ± 6.0

majority of the XRD data points, whereas the MLP algorithm appears to exhibit better fit with the XRD analysis data points within the middle Buntsandstein (Figure 13ii), particularly when compared to the other methods. This investigation highlights the need of applying a wide variety of techniques, each of which contributes to our comprehension of the mineral volume composition that is present within the Buntsandstein formations. Therefore, based on the quantitative results and qualitative description of the bulk mineral volume estimates within the Buntsandstein and Muschelkalk in this study, ML algorithms could be reliable alternatives in the absence of core samples for XRD analysis and generating a continuous description of mineral volumes for formations along borehole intervals.

Notwithstanding the promising capabilities of ML algorithms, it is essential to acknowledge their inherent limitations. Structural and mineralogical studies done on core and cutting samples within the Buntsandstein formations indicate the presence of many natural fractures filled with secondary hydrothermal minerals like barite ($BaSO_4$) in EPS-1 and anhydrite in GRT-1 (Genter et al., 1997; Vidal et al., 2018). For example, in EPS-1, a fracture sealed with a 5-cm-thick barite was observed around 1,205 m deep. These sealed fractures, filled with secondary minerals like barite and anhydrite within the Buntsandstein, elude detection by ML approaches, leading to the non-identification of secondary minerals. To overcome this limitation and ensure comprehensive mineral detection, we strongly recommend incorporating a broader spectrum of data sources, including petrophysical log data, as well as XRD and X-ray fluorescence (XRF) analyses as inputs for the ML approach.

5.2. Recommendations

We have demonstrated the potential of both statistical and ML approaches for accurate mineral volume estimation. Optimally, their validation relies on correlations with XRD analyses for bulk mineral volume estimation from core samples. Our example for the simple case of the Buntsandstein sandstones shows that they yield satisfactory estimates of mineral volume distributions. We point out the performance of ML algorithms that have exhibited significant effectiveness. However, like for any sophisticated methodology, several steps must be taken when applying them to data sets.

Several ML algorithms were applied in our study and a meticulous assessment was conducted to identify the strengths and weaknesses inherent in each method when applied to different minerals within the siliciclastic and carbonates. The analysis revealed variations in the performance of the algorithms: some provide realistic volume predictions for specific minerals while others do not. For instance, the prediction accuracy for Quartz was notably higher within the Buntsandstein formations (refer to Table 5) compared to the Muschelkalk formations (refer to Table 4). This discrepancy can be attributed to the substantial variance in training set sizes, as the simple, two-mineral models of the Buntsandstein were not as penalized for lack of an extensive data set. Consequently, we strongly recommend the utilization of a substantial number of data sets from multiple wells to improve the overall model accuracy. Furthermore, we assert that the accuracy of predictions for Muschelkalk and Buntsandstein formations would have markedly improved with a more comprehensive data set representative of the study area, surpassing the four boreholes considered in our analysis. It is therefore essential to tailor the selection of a ML algorithm to the specific characteristics of the data set and the geological formations under investigation.

To comprehend the prediction capabilities of each ML method, quantitative results should be combined with qualitative evidence from cores, cuttings, drilling masterlogs, and expert insights from existing literature. Additionally, based on the results of the parametric analysis discussed above, we strongly recommend a thorough analysis of each ML algorithm for mineral volume estimation.

6. Conclusion

Estimating mineral volume quantities in the URG infill, which consists of carbonates and siliciclastic, is a complex task due to the heterogeneous geological framework. This difficulty is also characterized by both linear and nonlinear challenges associated with geophysical data relationships. To tackle this inherent complexity, our methodology carefully combines the strong insights from traditional statistical methodologies with the latest breakthroughs in ML and AI. Our analysis focuses on the Muschelkalk and Buntsandstein formations, which are known to have a significant influence on geothermal fluid circulation. We employed three separate ML algorithms and to improve the accuracy of our model, we developed a novel hybrid model that utilizes a weighted average by concurrently integrating two distinct methods. The results of our investigation demonstrate a strong and reliable ability to predict outcomes using all of the methods employed. Our findings unveil a robust predictive performance across all methodologies used in this study. With strong consistency with qualitative mineral description from cores, cuttings, and expert geological knowledge of Muschelkalk and Buntsandstein formations. Lastly, The validity of the volume estimate methodologies must be assessed by a comparison with quantitative (e.g., XRD analysis of cores and cuttings, XRF estimates) or qualitative (e.g., core descriptions, XRD analysis, field observations, etc.) information.

Data Availability Statement

The research data utilized in this study have been supplied by ES-Geothermie and are subject to specific restrictions, encompassing nondisclosure agreements, licensing conditions, and proprietary constraints, rendering them unavailable to the general public. Should there be a desire to access this data, a formal request can be made to ES-Geothermie via geothermie@es.fr. It is important to note that the methods employed in this research are designed for reproducibility using comparable data from other geological locations.

Acknowledgments
This research is part of the SIMGEO project funded by ADEME (French Agency for Ecological Transition) with partners ITES, ES, CGG, and BRGM. The authors acknowledge the GEIE EMC and ECOGI for their help. Thanks to Michael Heap, Patrick Baud, Clément Baujard and Eléonore Dalmais for fruitful discussions. We also thank ES-Géothermie for access to parts of its petrophysical database. Additionally, we express our gratitude to CGG for granting us access to its software.

References

Agemar, T., Schellschmidt, R., & Schulz, R. (2012). Subsurface temperature distribution in Germany. *Geothermics*, *44*, 65–77. https://doi.org/10.1016/j.geothermics.2012.07.002

Aggarwal, C. C. (2018). *Neural networks and deep learning* (1st ed.). Springer. https://doi.org/10.1007/978-3-319-94463-0

Aichholzer, C., Duringer, P., & Genter, A. (2019). Detailed descriptions of the lower-middle Triassic and Permian formations using cores and gamma-rays from the EPS-1 exploration geothermal borehole (Soultz-sous-Forêts, Upper Rhine Graben, France). *Geothermal Energy*, *7*(1), 34. https://doi.org/10.1186/s40517-019-0148-1

Aichholzer, C., Duringer, P., Orciani, S., & Genter, A. (2016). New stratigraphic interpretation of the Soultz-sous-Forêts 30-year-old geothermal wells calibrated on the recent one from Rittershoffen (Upper Rhine Graben, France). *Geothermal Energy*, *4*(1), 13. https://doi.org/10.1186/s40517-016-0055-7

Amari, S. (1967). A theory of adaptive pattern classifier. *IEEE Transactions*, *EC*(16), 279–307.

Amosu, A., & Sun, Y. (2018). MinInversion: A program for petrophysical composition analysis of geophysical well log data. *Geosciences*, *8*(2), 65. https://doi.org/10.3390/geosciences8020065

Bächler, D., Kohl, T., & Rybach, L. (2003). Impact of graben-parallel faults on hydrothermal convection—Rhine graben case study. *Physics and Chemistry of the Earth, Parts A/B/C*, *28*(9), 431–441. (Heat Flow and the Structure of the Lithosphere). https://doi.org/10.1016/S1474-7065(03)00063-9

Baillieux, P., Schill, E., Edel, J.-B., & Mauri, G. (2013). Localization of temperature anomalies in the upper rhine graben: Insights from geophysics and neotectonic activity. *International Geology Review*, *55*(14), 1744–1762. https://doi.org/10.1080/00206814.2013.794914

Baujard, C., Genter, A., Dalmais, E., Maurer, V., Hehn, R., Rosillette, R., et al. (2017). Hydrothermal characterization of wells GRT-1 and GRT-2 in Rittershoffen, France: Implications on the understanding of natural flow systems in the rhine graben. *Geothermics*, *65*, 255–268. https://doi.org/10.1016/j.geothermics.2016.11.001

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, *13*, 1063–1095.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brun, J., & Gutscher, M.-A., & dekorp-ecors teams. (1992). Deep crustal structure of the rhine graben from dekorp-ecors seismic reflection data: A summary. *Tectonophysics*, *208*(1), 139–147. (Geodynamics of rifting, volume 1 Case history studies on rifts: Europe and Asia). https://doi.org/10.1016/0040-1951(92)90340-C

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Darnet, M., Vedrine, S., Bretaudeau, F., Marc, S., Capar, L., Pwavodi, J., et al. (2023). Towards a multi-physics multi-scale approach of deep geothermal exploration. *The Fourth EAGE Global Energy Transition Conference and Exhibition*, *2023*(1), 1–5. https://doi.org/10.3997/2214-4609.202321051

Deng, T., Ambía, J., & Torres-Verdín, C. (2019). Fast Bayesian inversion method for the generalized petrophysical and compositional Interpretation of multiple well logs with uncertainty quantification (Volume Day 4 Tue, June 18, 2019). https://doi.org/10.30632/T60ALS-2019_FFFF

Doveton, J. H. (2014). Compositional analysis of mineralogy. In *Principles of mathematical petrophysics* (Vol. 9). Oxford University Press. https://doi.org/10.1093/oso/9780199978045.003.0009

Duringer, P., Aichholzer, C., Orciani, S., & Genter, A. (2019). The complete lithostratigraphic section of the geothermal wells in Rittershoffen (upper rhine graben, eastern France): A key for future geothermal wells. *Bulletin de la Societe Geologique de France*, *190*, 9. https://doi.org/10.1051/bsgf/2019012

Duwiquet, H., Guillou-Frottier, L., Arbaret, L., Bellanger, M., Guillon, T., & Heap, M. J. (2021). Crustal Fault zones (CFZ) as geothermal power systems: A preliminary 3D THM model constrained by a multidisciplinary approach. *Geofluids*, *2021*, 1–24. https://doi.org/10.1155/2021/8855632

Eberl, D. (2003). User guide to RockJock—A program for determining quantitative mineralogy from X-ray diffraction data. U.S. Geological Survey. https://doi.org/10.3133/ofr200378

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. (Nonlinear Methods and Data Mining). https://doi.org/10.1016/S0167-9473(01)00065-2

Genter, A., Christian, C., & P, M. (1997). Evaluation de la fracturation des réservoirs par forages: Comparaison entre les données de carottes et d'imagerie de paroi. *Revue de l'Institut Français du Petrole*, *52*(1), 49–64.

Genter, A., & Traineau, H. (1991). Geological survey of the HDR borehole EPS1, Soultz-sous-Forêts, Alsace-France: Part 1-field report. In *BRGM Orléans IMRG, 32433*.

Genter, A., & Traineau, H. (1992). Borehole EPS-1, Alsace, France: Preliminary geological results from granite core analyses for hot Dry rock research. *Scientific Drilling*, *3*, 205–214.

Genter, A., & Traineau, H. (1996). Analysis of macroscopic fractures in granite in the HDR geothermal well EPS-1, Soultz-sous-Forêts, France. *Journal of Volcanology and Geothermal Research*, *72*(121–42), 1–2. https://doi.org/10.1016/0377-0273(95)00070-4

Guillou-Frottier, L., Carré, C., Bourgine, B., Bouchot, V., & Genter, A. (2013). Structure of hydrothermal convection in the Upper Rhine Graben as inferred from corrected temperature data and basin-scale numerical models. *Journal of Volcanology and Geothermal Research*, *256*, 29–49. https://doi.org/10.1016/j.jvolgeores.2013.02.008

Heap, M. J., Kushnir, A. R. L., Gilg, H. A., Marie, E. S. V., Pauline, H., & Patrick, B. (2019). Petrophysical properties of the Muschelkalk from the Soultz-sous-Forêts geothermal site (France), an important lithostratigraphic unit for geothermal exploitation in the Upper Rhine Graben. *Geothermal Energy*, *7*(27), 27. https://doi.org/10.1186/s40517-019-0145-4

Heap, M. J., Kushnir, A. R. L., Gilg, H. A., Wadsworth, F. B., Thierry, R., & Patrick, B. (2017). Microstructural and petrophysical properties of the Permo-Triassic sandstones (Buntsandstein) from the Soultz-sous-Forêts geothermal site (France). *Geothermal Energy*, *5*(26), 26. https://doi.org/10.1186/s40517-017-0085-9

Hillier, S. (1999). Quantitative analysis of clay and other minerals in sandstones by X-ray powder diffraction (XRPD). In *Clay mineral cements in sandstones* (pp. 213–251).

Hosseini, M. (2018). Formation evaluation of a clastic gas reservoir: Presentation of a solution to a fundamentally difficult problem. *Journal of Geophysics and Engineering*, *15*(6), 2418–2432. https://doi.org/10.1088/1742-2140/aacee3

Hu, K., Liu, X., Chen, Z., & Grasby, S. E. (2023). Mineralogical characterization from geophysical well logs using a machine learning approach: Case study for the horn river basin, Canada. *Earth and Space Science*, *10*(12), e2023EA003084. https://doi.org/10.1029/2023EA003084

Illies, J. (1972). The Rhine graben rift system-plate tectonics and transform faulting. *Geophysical Surveys*, *1*(1), 27–60. https://doi.org/10.1007/BF01449550

Ivakhnenko, G. L. V. A. G. (1967). *Cybernetics and forecasting techniques*. American Elsevier Publication Co.

Kappelmeyer, O., Gérard, A., Schloemer, W., Ferrandes, R., Rummel, F., & Benderitter, Y. (1991). European HDR project at Soultz-sous-Forêts: General presentation. Retrieved from https://api.semanticscholar.org/CorpusID:131888161

Laalam, A., Boualam, A., Ouadi, H., Djezzar, S., Tomomewo, O., Mellal, I., et al. (2022). Application of machine learning for mineralogy prediction from well logs in the Bakken Petroleum system (Volume Day 1 Mon, October 03, 2022). https://doi.org/10.2118/210336-MS

Lee, J., & Lumley, D. E. (2023). Predicting shale mineralogical brittleness index from seismic and elastic property logs using interpretable deep learning. *Journal of Petroleum Science and Engineering*, *220*, 111231. https://doi.org/10.1016/j.petrol.2022.111231

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Michael, J. H., Marlène, V., Kushnir, A. R., Farquharson, J. I., Baud, P., & Reuschlé, T. (2019). Rock mass strength and elastic modulus of the Buntsandstein: An important lithostratigraphic unit for geothermal exploitation in the Upper Rhine Graben. *Geothermics*, *77*, 236–256. https://doi.org/10.1016/j.geothermics.2018.10.003

Mitchell, W. K., & Nelson, R. J. (1988). A practical approach to statistical log analysis (Volume All Days).

Munck, F., Walgenwitz, F., Maget, P., Sauer, K., & Tietze, R. (1979). Synthèse géothermique du Fossé rhénan Supérieur. In *Commission of the European communities, 20*.

Mustafa, A., Tariq, Z., Mahmoud, M., Radwan, A. E., Abdulraheem, A., & Abouelresh, M. O. (2022). Data-driven machine learning approach to predict mineralogy of organic-rich shales: An example from qusaiba shale, rub' al khali basin, Saudi Arabia. *Marine and Petroleum Geology*, *137*, 105495. https://doi.org/10.1016/j.marpetgeo.2021.105495

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, *51*(3), 406–413. https://doi.org/10.1017/S0305004100030401

Pribnow, D., & Schellschmidt, R. (2000). Thermal tracking of upper crustal fluid flow in the Rhine Graben. *Geophysical Research Letters*, *27*(13), 1957–1960. https://doi.org/10.1029/2000GL008494

Pwavodi, J. (2023). Integration of borehole geophysical data and core petrophysical properties to model hydrogeological properties of the Nankai subduction zone (Theses, Université Grenoble Alpes [2019–2023]). Retrieved from https://theses.hal.science/tel-04368498

Pwavodi, J., & Doan, M.-L. (2023). Direct assessment of the hydraulic structure of the plate boundary at the toe of the Nankai accretionary prism. *Geophysical Journal International*, *236*(2), 1125–1138. https://doi.org/10.1093/gji/ggad473

Pwavodi, J., Kelechi, I. N., Angalabiri, P., Emeremgini, S. C., & Oguadinma, V. O. (2023). Pore pressure prediction in offshore Niger delta using data-driven approach: Implications on drilling and reservoir quality. *Energy Geoscience*, *4*(3), 100194. https://doi.org/10.1016/j.engeos.2023.100194

Rodriguez, O. H., & Lopez Fernandez, J. M. (2010). A semiotic reflection on the didactics of the Chain rule. *The Mathematics Enthusiast*, *7*(2), 10–332. https://doi.org/10.54870/1551-3440.1191

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rotstein, Y., Edel, J.-B., Gabriel, G., Boulanger, D., Schaming, M., & Munschy, M. (2006). Insight into the structure of the upper rhine graben and its basement from a new compilation of Bouguer gravity. *Tectonophysics*, *425*(1), 55–70. https://doi.org/10.1016/j.tecto.2006.07.002

Savre, W. C. (1963). Determination of a more accurate porosity and mineral composition in complex lithologies with the use of the sonic, neutron and density surveys. *Journal of Petroleum Technology*, *15*(09), 945–959. https://doi.org/10.2118/617-PA

Siddiqi, M. H., Alsayat, A., Alhwaiti, Y., Azad, M., Alruwaili, M., Alanazi, S., et al. (2022). A precise medical imaging approach for brain MRI image classification. *Computational Intelligence and Neuroscience*, *2022*, 15. https://doi.org/10.1155/2022/6447769

Vidal, J., Patrier, P., Genter, A., Beaufort, D., Dezayes, C., Glaas, C., et al. (2018). Clay minerals related to the circulation of geothermal fluids in boreholes at Rittershoffen (Alsace, France). *Journal of Volcanology and Geothermal Research*, *349*, 192–204. https://doi.org/10.1016/j.jvolgeores.2017.10.019

Villemin, T., Alvarez, F., & Angelier, J. (1986). The Rhinegraben: Extension, subsidence and shoulder uplift. *Tectonophysics*, *128*(1), 47–59. https://doi.org/10.1016/0040-1951(86)90307-0

Villemin, T., & Bergerat, F. (1987). L'evolution structurale du fosse rhenan au cours du Cenozoique; un bilan de la deformation et des effets thermiques de l'extension. *Bulletin de la Societe Geologique de France*, *III*(2), 245–255. https://doi.org/10.2113/gssgfbull.III.2.245

Xiao, J., Song, Y., & Li, Y. (2023). Comparison of quantitative X-ray diffraction mineral analysis methods. *Minerals*, *13*(4), 566. https://doi.org/10.3390/min13040566

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the twentieth international conference on international conference on machine learning* (pp. 856–863). AAAI Press.

Zhao, H., Ning, Z., Zhao, T., Zhang, R., & Wang, Q. (2016). Effects of mineralogy on petrophysical properties and permeability estimation of the Upper Triassic Yanchang tight oil sandstones in Ordos Basin, Northern China. *Fuel*, *186*, 328–338. https://doi.org/10.1016/j.fuel.2016.08.096

Erratum

The originally published version of this article contained typographical errors. Equation 4 should appear as follows:

$$\begin{bmatrix} \rho_{calc} & \rho_{Clay} & \rho_{dol} & \rho_{anh} & \rho_{qtz} \\ Pe_{calc} & Pe_{Clay} & Pe_{dol} & Pe_{anh} & Pe_{qtz} \\ K_{calc} & K_{Clay} & K_{dol} & K_{anh} & K_{qtz} \\ Vp_{calc} & Vp_{Clay} & Vp_{dol} & Vp_{anh} & Vp_{qtz} \end{bmatrix} \begin{bmatrix} V_{calc} \\ V_{Clay} \\ V_{dol} \\ V_{anh} \\ V_{qtz} \end{bmatrix} = \begin{bmatrix} \rho_{\log} \\ Pe_{\log} \\ K_{\log} \\ Vp_{\log} \end{bmatrix}$$

Additionally, in Table 2, the ΔT parameter for the last four minerals in the "Symbol or acronym" column should be replaced with the parameter *K*. The errors have been corrected, and this may be considered the authoritative version of record.