Drilling Fluid Properties Prediction: A Machine Learning Approach to Automate Laboratory Experiments

Mohammad J. Aljubran
Saudi Aramco (Saudi Arabia)
Hussain I. AlBahrani (hussain.albahrani@aramco.com)
Saudi Aramco (Saudi Arabia)
Jothibasu Ramasamy
Saudi Aramco (Saudi Arabia)
Arturo Magana-Mora
Saudi Aramco (Saudi Arabia)

Research Article

Keywords:

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1407939/v1

Drilling Fluid Properties Prediction: A Machine Learning Approach to Automate Laboratory Experiments

Mohammad J. Aljubran[1,+], Hussain I. AlBahrani[1,*,+], Jothibasu Ramasamy[1], Arturo Magana-Mora[1]

[1]Drilling Technology Team, EXPEC Advanced Research Center, Saudi Aramco, Dhahran, 31311, SA.

[*]hussain.albahrani@aramco.com

[+]these authors contributed equally to this work

ABSTRACT

Providing access to underground energy resources such as oil, gas, and geothermal energy requires drilling through subterranean rock formations. Drilling fluids are commonly used to enable this drilling process by serving a variety of functions. Chief among these is circulating into the drilled wellbore to remove the produced rock cuttings. Other functions include exerting the hydrostatic pressure necessary to prevent the flow of underground fluids into the wellbore while drilling, minimizing the invasion of solids and undesired fluids into the wellbore rock, and ensuring the fluid remains flowable within the means of the pumps available on site. To assess the ability of a drilling fluid to serve these functions, it has to have certain rheological properties, which are conventionally measured using specialized equipment. Considering that the formulation and components of these drilling fluids can vary greatly for different scenarios, the process of preparing samples and measuring their rheological properties, which is usually performed on a daily basis on a drilling site, is unavoidably time-consuming, repetitive, and error-prone. Based on this, it is apparent that there is a need for a computational model that can accurately predict drilling fluids properties based on the proposed concentration of their components without the need for further laboratory testing. This study describes a novel methodology to train a machine learning model derived from over 6,878 drilling fluid formulations to successfully predict water-based drilling fluids properties with a resulting R^2 of 91.07±6.35%.

Introduction

Drilling into subterranean rock formations is a complex process employed by different industries to achieve different objectives. These drilling operations can be means to extract energy resources, such as the case in the oil/gas and the geothermal energy industries, or to reduce carbon dioxide pollution through underground sequestration. The common denominator in the overwhelming majority of drilling operations is the use of drilling fluids. Drilling fluids are circulated in and out of drilled wellbores using surface pumps as part of the drilling rig structure. The main roles of these fluids are: 1) cleaning the wellbore by removing the rock cuttings produced by the drilling action of the drill bit at the bottom of the well; 2) applying the hydrostatic pressure necessary to overbalance the rock pore pressure, which serves to prevent the flow of the dangerous underground fluids into the wellbore; 3) minimizing the damage caused to the water or hydrocarbon-bearing subterranean reservoirs by arresting the invasion of undesired fluids or solids, and 4) cooling the wellbore to ensure the stability of downhole equipment. Fig. 1 illustrates the direction of the drilling fluid flow and hydrostatic/formation pressures. The composition of drilling fluids can vary significantly based on a multitude of considerations. For example, the density of the fluid is determined based on the depth and pore pressure of the rock, whereas the choice of chemical components used to formulate the fluid are contingent on: 1) the nature and type of rock to be drilled,; 2) the specific nature of the drilling operation; and 3) the local environmental regulations. For these reasons, drilling operations involve three major activities: drilling, casing, and cementing. These activities are conducted to drill each section or formation (as defined in a drilling program). Fig. 1B shows an example of a basic casing program of four sections. Clearly, each section would require a specific drilling fluid formulation as it would be unfeasible to drill from the surface to the desired depth with only one type of drilling fluid while accounting for all the formation pressure and chemical interaction requirements.

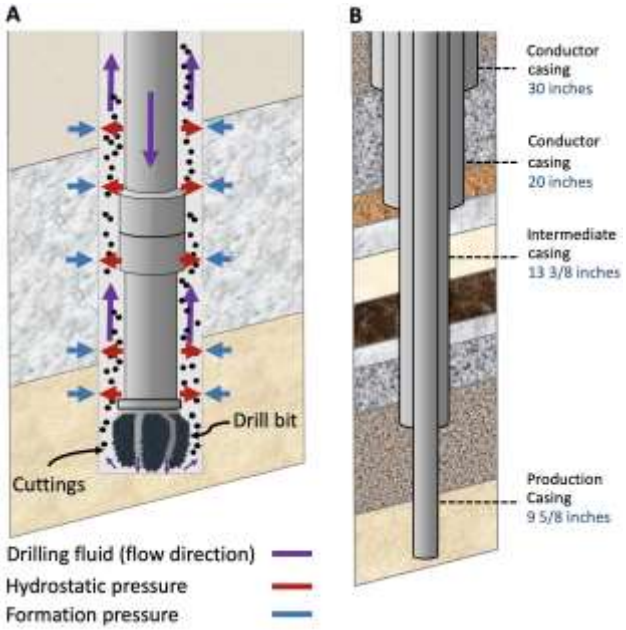

Figure 1. Illustration of the drilling process and casing program. A) The drilling fluid is circulated into the hole (through the drillstring) and back to the surface (through the annulus). B) Example of a casing program with four sections (30, 20, 13 3/8, and 9 5/8 inches).

Additionally, given that drilling environment and objectives can vary, there are three main types of drilling fluids to satisfy the requirements with these variations. These types are water-based drilling fluids or muds (WBM), oil-based or non-aqueous drilling fluids (NAF), and emulsion-based fluids. The consideration of these ever-changing variables means that possible combinations and concentrations of components or additives to formulate a drilling fluid are infinite. Consequently, this dictates that drilling fluids properties will always need to be measured in laboratories based on their proposed components and formulations as there is no available deterministic model for those properties. This is essential as these measurements are the only means to ensure that a drilling fluid will fully serve its aforementioned roles within the context and specifics of the planned drilling operation.

Overview of relevant drilling fluids properties

The following is a discussion of the correlations between the main drilling fluid properties and their functions. The properties covered in this discussion are: 1) the fluid density; 2) the rheological constants of the yield point (YP) and plastic viscosity (PV) expressed in centipoise (cP); 3) the gel strength; 4) the filtration fluid

loss; and 5) pH. The drilling fluid density, also known as the mud weight, is usually measured in pound-force per cubic foot (pcf). In conventional drilling, the fluid density is set at a value that ensures the hydrostatic pressure provided by the drilling fluid is higher than the rock pore pressure to maintain wellbore stability and avoid influxes of gas, oil, or water (see Fig. 1). To adjust the fluid density, a drilling fluid component, known as the weighting agent, is added into the formulation in a concentration that equates the fluid density to the target value. Basic volumetric and mass balance calculations based on the specific gravities of each drilling fluid component are performed to calculate the final overall density. The second drilling fluid property to be considered is the YP. This property is a parameter of a commonly used rheological model, which is the Bingham-plastic model [1]. Rheological models are frequently used to describe the flow behavior of drilling fluids. In this specific model, the shear stress (τ) is a linear function of the shear rate (γ). The YP is the y-axis intercept, as illustrated in Fig. 2, which signifies the threshold value of shear stress required to initiate the fluid flow. The slope of the linear line is the PV, essential for wellbore hydraulics and hole cleaning [2]. This model is expressed as follows:

$$\tau = YP + PV(\gamma) \tag{1}$$

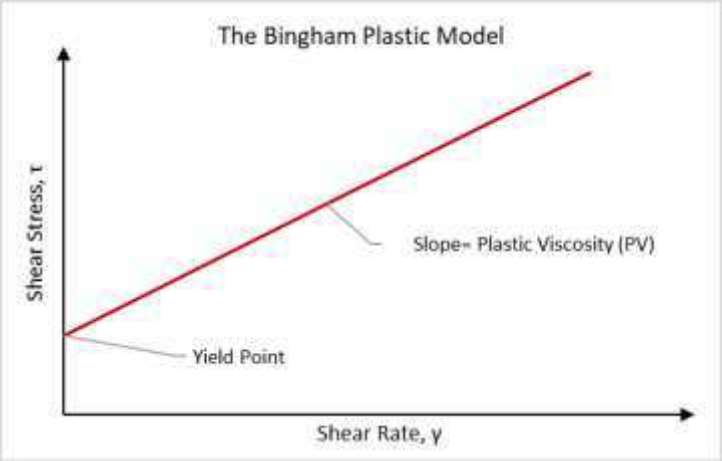

Figure 2. Graphical representation of the Bingham-Plastic rheology model.

The two parameters describing the Bingham-plastic model can be measured using a viscometer. In this device, a rotating outer cylinder at known revolutions per minute (RPM) values is used to assess the viscous drag exerted by the drilling fluid sample. This is achieved by measuring the torque created on an inner cylinder in the

equipment as a result of the viscous drag. By performing these measurements at 600 RPM and 300 RPM, PV and YP can be calculated as follows:

$$PV = \theta_{600\ RPM} - \theta_{300\ RPM} \quad (2)$$

$$YP = \theta_{300\ RPM} - PV \quad (3)$$

where $\theta_{600\ RPM}$ and $\theta_{300\ RPM}$ are the viscometer dial readings at 600 RPM and 300 RPM, respectively. The value for the YP is used to assess a drilling fluid formulation ability to carry the rock cuttings from the drilled wellbore. This value is essential for ensuring a proposed drilling fluid formulation is capable of cleaning the newly drilled wellbore section in an efficient manner that would allow drilling operations to progress smoothly. The gel strength is another relevant rheological property, which is also measured using a viscometer. This property is used to assess the force or shear stress required to initiate the flow of drilling fluids after a certain period of stagnation. This is measured in a viscometer by taking the dial readings at low shear rates ($\theta_{3\ RPM}$) after keeping the drilling fluid sample static for some time (often 10 seconds, 10 minutes, or 30 minutes). From an operational perspective, the gel strength property is essential for determining a drilling fluid's ability to keep rock cuttings and other solids suspended while the fluid is not flowing. The non-flow periods are periods where the surface pumps are shut-off to perform necessary and routine operations, such as extending the drill-pipe string to enable drilling to a deeper point. Collectively, the gel strength measurement is relied on to ensure that suspended solids during stagnation periods will not settle down into the bottom of the well and to ascertain that the pumps will be able to reinitiate flow.

With respect to the fourth considered drilling fluid property, the filtration fluid loss is used to evaluate seepage tendencies of a drilling fluid formulation. Any drilling fluid formulation will always consist of solid particles and fluids. When a fluid formulation is subjected to conditions mimicking those experienced while drilling in terms of pressure and temperature, the solid particles can aggregate on the wellbore wall rock and work to filter fluids out of the wellbore and into the rock formation. This process of solids aggregation and fluids seepage can have a severe damaging influence on the rock formation being drilled. The damage can eventually be exhibited in plugged rock pores, which can significantly jeopardize the objectives of the well being drilled. This property is measured in a laboratory by subjecting a drilling fluid sample to high pressure and temperature in an isolated cell and using a filtering element that allows solid particles to aggregate and seepage to take place. The volume of the fluid seeping through the flirting element is gathered and measured. This volume is designated as the filtrate volume. The thickness of the solid particles aggregation is also measured and defined as the mud cake thickness.

An evaluation of these two values is used to assess a proposed drilling fluid formulation ability to minimize the damage caused to the water or hydrocarbon-bearing subterranean reservoirs. It should also be noted that the measurement of these values can be performed either through a standard procedure set by the American Petroleum Institute (API) or by high pressure and high temperature (HPHT) equipment.

The final property considered is the pH of the drilling fluid formulation. The value on the pH scale is used to evaluate two main chemical interaction concerns. The first concern is that of corrosivity as acidic formulations, i.e., those with a pH value below seven, will have a damaging effect on tubulars and equipment exposed to the drilling fluid. The second concern is that of general chemical interaction between the drilling fluid formulation and the formation rock. For example, in cases where the drilled formation rock type is carbonate, acidic fluid formulations will need to be avoided as they can trigger a chemical interaction that damages the integrity of the wellbore rock.

The need for a predictive model

The standard practice for measuring the relevant drilling fluid properties is to initially evaluate a proposed formulation by mixing a sample of the fluid and measuring its properties in a lab. Once the fluid formulation is decided and the drilling operation commences, samples are taken from the site of the drilling rig fluid tanks on a daily basis for confirmation testing to ensure the stability of the fluid properties throughout the duration of the drilling operation. This confirmation testing is essential when the fluid formulation is adjusted in response to certain events that take place while drilling. For instance, when the rock pore pressure is higher than the anticipated value, the drilling fluid density will need to be increased. This change in density can lead to a significant shift in all the other properties. Therefore, it is necessary to repeat these measurements at all times. Based on this, it can be seen that measuring the drilling fluid properties is a time-consuming, repetitive, and error-prone process.

While possible combinations and concentrations of components to formulate a drilling fluid are infinite, the overall drilling operations in well-developed fields can utilize a wide variety of formulations, which can be representative of all possible combinations. Given the accessibility to such a varied and extensive dataset along with the apparent need to reduce the reliance on repetitive and time-consuming lab tests, building a machine learning (ML) model based on the lab results of a large number of drilling fluid formulations can serve to enable the prediction of a fluid

formulation properties based on its components and their concentration. Deriving a robust and accurate ML model is the objective of this paper. This work relies on a dataset that contains 6,878 water-based drilling fluid formulations along with lab measurements of relevant properties.

Related work

Although several studies using computational models have been proposed for the prediction and detection of multiple drilling applications [3-6], only a few studies have focused on the applications of ML in drilling fluids analysis. Among these studies, Gowida et al. [7] proposed a data-driven framework to predict the rheological properties of a specific drilling fluid formulations type, which is the CaCl2 brine-based drill-in fluid. The term drill-in fluid refers to formulations that are specifically designed for drilling through hydrocarbon-bearing reservoirs as they can ensure minimizing solids content and formation rock damage. This reference relies on measurements of the drilling fluid density, PV, YP, and a property known as the Marsh funnel viscosity, which is performed on-site in the drilling rig. Within their proposed framework, artificial neural networks (ANN) are used to derive a supervised learning model to predict the properties mentioned above and calculate more relevant rheological properties based on the predicted values. There are six main fundamental differences between the framework proposed by Gowida et al. and the work detailed in this study. First, data quality concerns are alleviated by relying on lab-produced measurements rather than field-based. This is a crucial aspect to consider as drilling rig or field-based measurements of drilling fluid properties notoriously suffer from inaccuracies and data quality problems due to the lack of sophisticated equipment and the necessary environment for the required analysis. Second, the model in this reference utilizes the drilling fluid density and the Marsh funnel viscosity as the input features rather than using the drilling formulation components along with their concentrations, which is the methodology employed in our work. Relying solely on these two features as the input parameters is only possible because the third fundamental difference between this reference and the work presented in this study is the lack of variations in terms of drilling formulations. As mentioned previously, this reference aims to predict the properties of a CaCl2 brine-based drill-in fluid, which is only a specialized type of drilling fluids formulations that falls within the water-based drilling fluids group. Relying on the specific components of the formulation along with their concentration as input features allows for a more robust and versatile model that can be generalized across a wide range of applications. This is also a more time-efficient scheme as it does not

require initial lab testing results in order to predict the desired properties. The fourth fundamental difference is the dataset size and the variations within it. The dataset used in this study consists of 6,878 data points encompassing a wide range of formulations and their lab-measured properties. In the reference in question, there are 515 data points, which are based on field-based measurements performed at a frequency of once every 15-20 min. This frequency of measurements and the size of the dataset allude to the lack of variation in the dataset. This lack of variation is especially suspect when evaluating the performance of the ML model through a randomized data split.

Theron et al. [8] proposed an ML model derived from a specialized type of drilling fluids. This reference aims to predict the rheological behavior of spacer fluids under the influence of a wide range of temperatures. Spacer fluids are used in drilling operations to provide a buffer between the drilling fluid and a cement slurry pumped into the well [9]. Based on the specialized role played by this fluid, its functions and properties differ from those of drilling fluids. For example, spacer fluids need to efficiently displace drilling fluids to make way for a cement slurry to be pumped; therefore, spacer fluid properties are contingent on those of drilling fluid and the cement slurry. In the framework presented in this reference, the authors considered the temperature, spacer fluid components and their concentration, the spacer fluid density, and the corresponding shear rate (γ) as the input features to derive an ANN. The output of the model is the shear stress (τ), which, together with the input shear rate, can be used to calculate the desired rheological properties, as illustrated in Fig. 2.

Some references suggest the integration of an alternative form of on-site measurements or sensing along with ML methods to predict drilling fluid properties. An example of this is the use of the sonic properties of drilling fluids through the use of ultrasonic-through-transmission [10]. In this reference, a methodology for real-time estimation of drilling fluids properties is presented by utilizing ultrasonic transducers. The measurements were performed on a total of 22 water-based drilling fluid samples, and the signal information along to predict density, PV, and gel strength. While such a framework is beneficial in producing real-time measurements for monitoring the drilling fluid behavior and influences, unlike the scheme we are presenting, it does not contribute to the process of formulating a new drilling fluid recipe based on changes in the operational objectives.

Another reference employs ML models to predict the consequential behavior of a drilling fluid formulation based on its rheological properties rather than predict the

properties themselves [11]. In this reference, the rheological properties based on the Herschel-Bulkley rheology model, which is an alternative to the Bingham-plastic model considered in our work, are used among other features to predict the pressure drop in the annular space between the drill pipe and the wellbore wall. While the application of the ML model is different in terms of objectives and methodology, this reference does illustrate the value and need for accurate drilling fluid properties as they can lead to accurate pressure drop calculations, which in turn will lead to more efficient hole cleaning capabilities and avoidance of drilling troubles such as stuck pipe incidents.

Results

The main contribution of the study is the description of the novel approach that uses drilling mud formulations to derive an ML model to estimate drilling fluid properties.

We performed a systematic comparison of ML models and compared their ability to accurately predict the minimum and maximum expected values for each of the fluid properties of interest. We derived an independent ML model for each of the fluid properties due to the multi-output regression model challenges mentioned in the Methods Section. We used MAE to contrast and rank the tuned LR, ANN, and ERT models as this metric is simple, interpretable, and robust to outliers when compared to MSE and R2. Note that it is unusual to have outliers in this application, so they should not be emphasized during the ML model development process.

We split the fluid formulations dataset into two groups: 80% training and 20% for holdout testing, which represent 5,503 and 1,375 fluid formulations, respectively. We used a three-fold cross-validation technique within the training data split (80% of the data) to tune model hyperparameters.

We compared the results on the testing set (20% of data) obtained by the tuned LR, ANN, and ERT models, shown in Table 1. Notably, the results indicated that tree-based algorithms were superior compared to the other models, and ERT consistently outperformed the other models for the prediction of all considered fluid properties. This may be because the dataset is characterized by complex non-linear relationships, high-dimensional input space, missing data, and sparsity. The input matrix (representing the 780 drilling additives) has a sparsity of 97.59%, measured as the ratio of zero elements to the total number of elements [12]. The ERT algorithm is similar to the random forest algorithm in terms of randomly choosing subsamples of the input features when developing each single decision tree. While random forests bootstrap the data points and perform optimal split at

each node using an entropy metric, the ERT algorithm does not bootstrap data samples, nor does it perform optimal splitting at nodes. Instead, ERT performs random splits within the range of each feature. Whereas optimal splitting is biased by sparsity, random splitting may produce more robust models. In addition, random splitting often yields more leaves or terminal nodes across the ERT decision trees, which generally leads to low-variance models compared to random forest and single decision tree algorithms [13].

Fluid Property (unit)	LR (MAE)	ANN (MAE)	ERT (MAE)
YP ($lb/100ft^2$)	1.12	0.75	**0.26**
PV (cP)	1.88	1.72	**0.70**
10-min Gel Strength ($lb/100ft^2$)	2.05	0.99	**0.42**
10-sec Gel Strength ($lb/100ft^2$)	1.53	0.79	**0.27**
Filtrate Volume (ml)	0.85	0.51	**0.10**
API Fluid Loss (ml)	0.48	0.30	**0.10**
pH	0.26	0.15	**0.05**
Average	1.17	0.74	**0.27**

Table 1. MAE comparison of models. Results highlighted in bold font indicate the best performing results

Discussion

Table 2 shows the optimal ERT hyperparameters, MAE, and R2 results of the best cross-validated models on the holdout test set for each fluid property. Note that "min_samples_split" is frequently optimal at a value of two, which indicates that trees tend to perform better when splitting non-terminal nodes until reaching terminal nodes (leaves) with very few samples. Meanwhile, "n_estimators" and "max_features" varied across the range of random search and demonstrated that fluid properties vary in complexity and nonlinearity. The ERT achieves accurate prediction results on the holdout test set with R2 of 91.07±6.35%.

Among the multiple applications for the fluid property prediction model is the fluid reformulation that can be done by modifying the additive concentrations to optimize costs or any other criteria. Additionally, as an analogy to the computational model for high-throughput screening of chemical compounds used in drug repurposing [14], the fluid property prediction model can be used as a tool to optimize additive concentrations while accounting for toxicity, costs, availability, among other constraints for each additive. From the performed experimental results (not shown), we were observed that while ANNs models were outperformed by ERT, ANNs may still be a better option when trying to create an entirely new fluid formulation. This is because ERT models may produce optimistic results as they are purely based on a split threshold. In other words, an ERT model may not be able to correctly account for the effects of a particular additive with an extreme concentration simply because that concentration is greater than the threshold. Alternatively, an ANN will be considerably more affected by an additive at a very high concentration. Nevertheless, ERT models remain the preferred models when testing formulations derived with some prior knowledge (e.g., small concentration perturbations on an already known fluid).

Mud Property (unit)	Optimal Hyperparameter			MAE	R^2
	n_estimators	max_features	min_samples_split		
YP ($lb/100ft^2$)	94	0.4	2	0.26	0.9019
PV (cP)	73	0.5	2	0.70	0.7751
10-min Gel Strength ($lb/100ft^2$)	178	0.9	16	0.42	0.9525
10-sec Gel Strength ($lb/100ft^2$)	178	0.8	16	0.27	0.9675
Filtrate Volume (ml)	94	0.4	2	0.05	0.9750
API Fluid Loss (ml)	94	0.4	2	0.10	0.9171
pH	9	0.4	2	0.05	0.8855
			Average	0.26	0.9107

Table 2. ERT models results on the holdout test set and their respective hyperparameters.

In this study, we presented a novel approach to derive an ML model for the prediction of drilling fluid properties.

The presented model uses the drilling fluid additives and their corresponding concentrations as features. This approach enabled the model to predict a group of parameters referring to the drilling fluid properties that are most relevant to the efficiency of drilling operations. Additionally, rather than narrow the prediction space to a specific drilling fluid formulation, as is the case in related models described in the literature, the presented model encompasses a wide range of water-based fluid formulations. This level of generalization is achieved through the dataset, which contains 780 unique additives. A total of 6,878 water-based fluid formulations were used for model training and testing. From the systematic performance comparison of 26 different ML algorithms, the ERT models consistently outperformed other models and were able to predict fluid properties with an average R^2 of 91.07±6.35%.

The presented model is beneficial for the experts designing drilling fluid formulations in a lab environment for the purpose of issuing recommendations for the drilling site. This is because the properties for each candidate drilling fluid formulation can be estimated with minimal lab tests. As a result, the routine, time-consuming, repetitive, and error-prone laboratory work necessary for designing a formulation with specific required properties can be avoided or greatly minimized. Therefore, the consumption of chemicals and labor for optimizing a formulation could be significantly minimized by applying the model. As the presented model focuses on optimizing the drilling fluid design process in a lab environment, it also provides a pathway towards developing models for direct field or drilling operations applications. The proposed setup of the model allows for considering the impact of the interaction between the drilling fluid and the subterranean rock formations while drilling. The influence of this interaction can potentially be reflected through the incorporation of the concentration of rock cuttings as a separate input feature into the ML model. Another potentiality created by the presented model is reducing expenditure from drilling fluid chemicals consumption. This can be achieved by optimizing the concentration and cost of available drilling fluid additives against the targeted fluid properties.

Methods

The proposed approach in this study aims to leverage the vast amount of collected drilling fluid recipes to derive robust ML models able to estimate the multiple rheology properties. The following subsections introduce the data collection and preprocessing steps, model training, and evaluation criteria to assess the performance of the models.

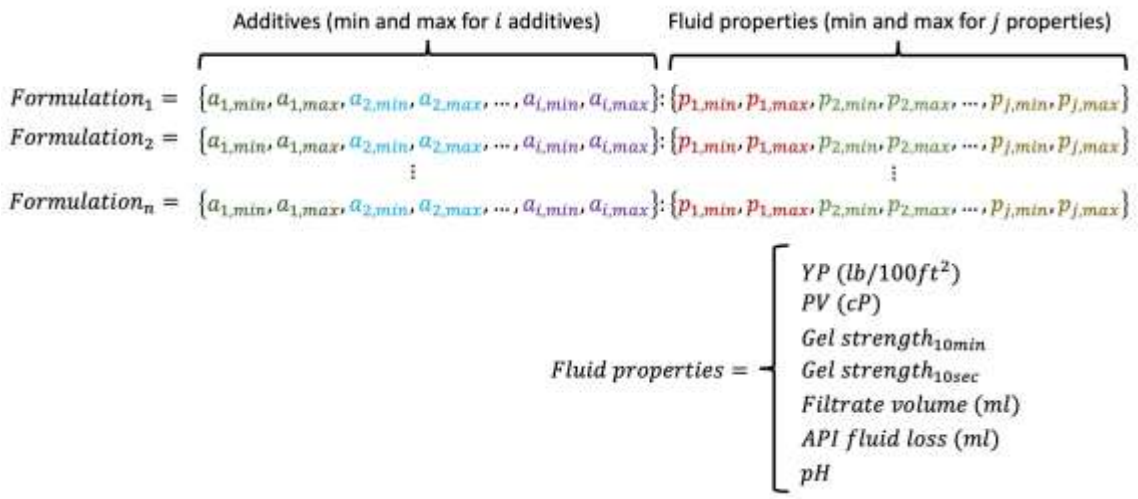

Figure 3. Drilling fluid formulations. Each formulation is defined by the minimum and maximum concentrations of *i* additives and the minimum and maximum values for the expected *j* fluid properties. The dataset contains 6,878 water-based formulations spanning 780 additives for 15 different hole sections.

Data collection and preprocessing

Data quality and scarcity problems represent a significant challenge for developing robust ML models, often the case when using experimental data that rely on manual laboratory testing. As different wellbore sections and fields (characterized by different lithology) require a completely different fluid formulation with different additives, the resulting fluid recipe formulation may be characterized as a sparse matrix. If the objective is to derive a model able to account for all sections and fields, then the dataset must contain a sufficiently large and diverse set of fluid recipes with enough variations to define the resulting effects of each additive on the fluid properties.

Consequently, the dataset used in this study consists of a large and diverse amount of fluid formulations. Each fluid formulation is defined by a set of compositional additive concentrations (e.g., water, bentonite, starch, etc., often measured in barrels or pounds per barrel) and their respective fluid properties (YP, PV, 10-min gel strength, 10-sec gel strength, filtrate volume, and pH). In this study, we considered a total of 11,174 water-based fluid formulations used in three oil fields for 15 different hole sections (from 28 inches to 3 5/8 inches), spanning a total of 780 unique additives. Because of the complex interactions and effects of the additive concentrations on the fluid properties, the dataset describes an acceptable

concentration range for each additive (minimum and maximum) and a range for the expected fluid properties (Fig. 3).

We performed the following data preprocessing steps to ensure the fluid formulations contained all relevant information. Firstly, we removed formulations where both maximum and minimum fluid properties are missing. If a record/formulation retains either the minimum or maximum of a property, then minimum and maximum are set to be equal, and the record is maintained. In addition, each property cannot be greater or smaller than specific thresholds based on the standard engineering design and procedural requirements across the operations for which these fluid recipes were initially formulated. Hence, records exceeding these thresholds are deemed incorrect, likely due to human error during the data input process. These preprocessing operations resulted in a dataset containing 6,878 records/formulations. Table 3 summarizes the thresholds and the final number of records for each property after preprocessing.

Fluid Property (unit)	Threshold		Formulation Count
	Minimum	Maximum	
YP ($lb/100ft^2$)	15	40	5,802
PV (cP)	0	60	1,149
10-min Gel Strength ($lb/100ft^2$)	2	50	3,628
10-sec Gel Strength ($lb/100ft^2$)	1	30	5,113
Filtrate Volume (ml)	0	30	1,954
API Fluid Loss (ml)	0	15	2,766
pH	7	12	6,127

Table 3. MAE comparison of models. Results highlighted in bold font indicate the best performing results

Machine learning model

The application of artificial intelligence algorithms has achieved significant milestones over the past few years. Particularly, with the increase of computational resources and amount of available data, ML algorithms have developed rapidly and been adopted widely across various domains and industries [15-18]. Mitchell [19] defines an ML algorithm as a process where "a computer program is said to learn

from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E". In this study, we aim to develop ML models to perform multi-output regression. The referenced task T aims to predict seven fluid properties: YP, PV, 10-min gel strength, 10-sec gel strength, filtrate loss, API fluid loss, and pH. The fluid formulations mixed in the laboratory and their corresponding property measurements represent the experience E, which the ML model aims to learn. The basis of the learning process is to obtain a functional form to map the inputs (fluid additives) to the desired outputs (fluid properties). Since all outputs are continuous variables, the performance P can be measured using the standard regression metrics, such as mean absolute error (MAE), square error (MSE), coefficient of determination (R2), amongst others [20].

Fluid property prediction is a multi-output regression task involving more than one output variable given one or more input variables. Since fluid properties are physically and chemically correlated, it is fair to first consider multi-output regression where the model learns a common functional form to represent all properties at once. However, multi-output learning is associated with multiple challenges that can be described using the four Vs. paradigm (i.e., volume, velocity, variety, and veracity) [20, 21]. Multi-output regression is associated with data volume challenges where the output dimension considerably expands, hindering the ML model learning process. It also leads to data imbalance where different outputs have different records count. This is also related to data velocity, where the time and cost required to acquire measurements for different outputs are dependent on the complexity of the corresponding laboratory equipment and test procedure. Data variety is the third component of the four Vs. paradigm, where outputs are different in nature and scale, which poses challenges in formulating a balanced loss function, which correctly prevents biased learning that could favor higher prediction accuracy in one output over the other. Lastly, veracity is another challenge in multi-output regression where different outputs lead to different levels of noise, missing records, incompleteness, among others. Because of these reasons, we derived a separate model for the prediction of each fluid property. Note that we aim to predict the potential minimum and maximum values (outputs) that each property could take given ranges of fluid additives (inputs), as shown in Fig. 3. Hence, we developed seven binary-output regression models, where each is tasked to predict the maximum and minimum of each fluid property. The number of records/fluid formulations for each property/model is described in Table 3.

The selection of the ML algorithm to derive the best-performing model for a given problem is not a trivial task as the performance of these models depends on the size

of data, input-input, and input-output correlation complexity [22-26]. Consequently, there exist many ML algorithms that vary in complexity, robustness, and interpretability that may need to be empirically validated and tested [27]. To identify the best performing model for this particular regression problem, we used the PyCaret framework [28] with default model parameters to evaluate the performance of 25 different ML regression algorithms from multiple families, such as regression (e.g., polynomial regression), instance-based (e.g., k-nearest neighbors and support vector machines), regularization (e.g., lasso regression - LR), tree-based (e.g., regression trees), Bayesian (e.g., naïve Bayes), dimensionality reduction (e.g., linear discriminant analysis and quadratic discriminant analysis), and ensemble algorithms (e.g., random forest, extremely randomized trees - ERT, boosting, Ada-boosting, gradient boosting machines).

From the comparison of these models, we considered the ML algorithm that produced the best performing results, i.e., ERT [29], and used the Scikit-learn [30] implementation and a random search algorithm to tune the hyperparameters (shown in Table 4). In addition to the ERT algorithm, we considered the ANN [31] and LR [32] Scikit-learn implementations for further analysis and tuned the hyperparameters (shown in Table 4). Although LR did not achieve the best results, we included it for further discussion.

Hyperparameter (Algorithm)	Description	Range
alpha (LR)	multiplier of L1 regularization in the LR algorithm loss function	[0.0, 1.0]
alpha (NN)	multiplier of L2 regularization in the NN algorithm loss function	[0.0, 1.0]
learning_rate (NN)	learning rate used in the Adam optimizer of NN models	[1E-4, 1.0]
Activation (NN)	Activation function for the hidden layers of NN models	['relu', 'sigmoid', 'tanh']
hidden_layer_count (ANN)	Number of hidden layers used in NN models	[1,2]
hidden_layer_sizes (ANN)	Number of neurons in each hidden layer of NN models	[1, 100]
n_estimators (ERT)	integer representing the number of tree estimators within a regression tree	[10, 200]
max_features (ERT)	float representing the percentage of total features to be considered when searching for the best split after randomly generating thresholds	[0.1, 1.0]
min_samples_split (ERT)	integer representing the minimum number of samples required to split an internal node in a tree	[2, 100]

Table 4. Model hyperparameters search space considered in a 100 iterations random search algorithm

References

[1] Bingham EC. An investigation of the laws of plastic flow: US Government Printing Office, 1917.

[2] Gonzalez M, Thiel T, Gooneratne C, Adams R, Powell C, Magana-Mora A, et al. Development of an In-Tank Tuning Fork Resonator for Automated Viscosity/Density Measurements of Drilling Fluids. IEEE Access. 2021;9:25703-15.

[3] Aljubran M, Ramasamy J, Bassam M, Magana-Mora A. Deep Learning and Time-Series Analysis for the Early Detection of Lost Circulation Incidents During Drilling Operations. IEEE Access. 2021.

[4] Magana-Mora A, Abughaban M, Ali A. Machine-Learning Model for the Prediction of Lithology Porosity from Surface Drilling Parameters. Conference Machine-Learning Model for the Prediction of Lithology Porosity from Surface Drilling Parameters. Society of Petroleum Engineers.

[5] Gooneratne CP, Magana-Mora A, Otalvora WC, Affleck M, Singh P, Zhan GD, et al. Drilling in the Fourth Industrial Revolution—Vision and Challenges. IEEE Engineering Management Review. 2020;48(4):144-59.

[6] Magana-Mora A, Gharbi S, Alshaikh A, Al-Yami A. AccuPipePred: A framework for the accurate and early detection of stuck pipe for real-time drilling operations. Conference AccuPipePred: A framework for the accurate and early detection of stuck pipe for real-time drilling operations. Society of Petroleum Engineers.

[7] Gowida A, Elkatatny S, Ramadan E, Abdulraheem A. Data-Driven Framework to Predict the Rheological Properties of CaCl2 Brine-Based Drill-in Fluid Using Artificial Neural Network. Energies. 2019;12(10):1880.

[8] Theron B, Bodin D, Fleming J. Optimization of spacer rheology using neural network technology. Conference Optimization of spacer rheology using neural network technology. Society of Petroleum Engineers.

[9] Carney L. Cement spacer fluid. Journal of Petroleum Technology. 1974;26(08):856-8.

[10] Jondahl MH, Viumdal H. Estimating Rheological Properties of Non-Newtonian Drilling Fluids using Ultrasonic-Through-Transmission combined with Machine Learning Methods. Conference Estimating Rheological Properties of Non-Newtonian Drilling Fluids using Ultrasonic-Through-Transmission combined with Machine Learning Methods. IEEE, p. 1-4.

[11] Kumar A, Ridha S, Ganet T, Vasant P, Ilyas SU. Machine Learning Methods for Herschel–Bulkley Fluids in Annulus: Pressure Drop Predictions and Algorithm Performance Evaluation. Applied Sciences. 2020;10(7):2588.

[12] Hurley N, Rickard S. Comparing measures of sparsity. IEEE Transactions on Information Theory. 2009;55(10):4723-41.

[13] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006;63(1):3-42.

[14] Soufan O, Ba-Alawi W, Magana-Mora A, Essack M, Bajic VB. DPubChem: a web tool for QSAR modeling and high-throughput virtual screening. Scientific reports. 2018;8(1):1-10.

[15] Aljubran MJ, Horne R. Prediction of Multilateral Inflow Control Valve Flow Performance Using Machine Learning. SPE Production & Operations. 2020.

[16] Albaradei S, Magana-Mora A, Thafar M, Uludag M, Bajic VB, Gojobori T, et al. Splice2Deep: An ensemble of deep convolutional neural networks for improved splice site prediction in genomic DNA. Gene: X. 2020;5:100035.

[17] AlBahrani H, Papamichos E, Morita N. Building an Integrated Drilling Geomechanics Model Using a Machine-Learning-Assisted Poro-Elasto-Plastic Finite Element Method. SPE Journal. 2021:1-21.

[18] Albalawi F, Chahid A, Guo X, Albaradei S, Magana-Mora A, Jankovic BR, et al. Hybrid model for efficient prediction of poly (A) signals in human genomic DNA. Methods. 2019;166:31-9.

[19] Mitchell TM. Artificial neural networks. Machine learning. 1997;45:81-127.

[20] Xu D, Shi Y, Tsang IW, Ong Y-S, Gong C, Shen X. Survey on multi-output learning. IEEE transactions on neural networks and learning systems. 2019;31(7):2409-29.

[21] Borchani H, Varando G, Bielza C, Larranaga P. A survey on multi‐output regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2015;5(5):216-33.

[22] Cho G, Yim J, Choi Y, Ko J, Lee S-H. Review of machine learning algorithms for diagnosing mental illness. Psychiatry investigation. 2019;16(4):262.

[23] Murugan NS, Devi GU. Detecting spams in social networks using ML algorithms-a review. International Journal of Environment and Waste Management. 2018;21(1):22-36.

[24] Pineda-Jaramillo JD. A review of Machine Learning (ML) algorithms used for modeling travel mode choice. Dyna. 2019;86(211):32-41.

[25] Crisci C, Ghattas B, Perera G. A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling. 2012;240:113-22.

[26] Magana-Mora A, Bajic VB. OmniGA: Optimized omnivariate decision trees for generalizable classification models. Scientific Reports. 2017;7(1):1-11.

[27] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research. 2014;15(1):3133-81.

[28] Gain U, Hotti V. Low-code AutoML-augmented Data Pipeline–A Review and Experiments. Conference Low-code AutoML-augmented Data Pipeline–A Review and Experiments, vol. 1828. IOP Publishing, p. 012015.

[29] Loh WY. Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery. 2011;1(1):14-23.

[30] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

[31] Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks. Chemometrics and intelligent laboratory systems. 1997;39(1):43-62.

[32] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

Authors contribution statement

M.J.J. designed and developed the computational models, analyzed the data, and contributed to the manuscript write up. H.I.B. conceptualized and designed the study and contributed to the manuscript write up. J.R. performed the necessary lab experiments and contributed to the manuscript write up. A.M.M extracted, processed, and analyzed the dataset and contributed to the design of the computational models and manuscript write up. All authors reviewed the final version of the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.