Jurnal Ilmu Fisika
Vol. 18, No. 1, March 2026, pp. 80–96
ISSN: 1979-4657 (Print); 2614-7386 (Online)
https://doi.org/10.25077/jif.18.1.80-92.2026

Jurnal Ilmu Fisika
Open Access

A Computational Physics–Based Machine Learning Modelling of Multiphase Flow Dynamics for Crude Oil Percentage Prediction Using Water Cut and Sediment Indicators

Jesi Pebralia[1], Iful Amri[2], Dwi Rahmah Amanda[1], Muhammad Aziz Kurniawan[3]

[1] Physics Study Program, Faculty of Science and Technology, Universitas Jambi, Muaro Jambi, 36361, Indonesia
[2] Physics Education Study Program, Faculty of Teacher Training and Education, Universitas Sriwijaya,
Ogan Ilir, 30128 Indonesia
[3]PT. Pertamina EP Field 1 Jambi, Jambi, 36129, Indonesia

Article Info

Article History:

Received December 04, 2025
Revised February 21, 2026
Accepted February 25, 2026
Published online March 01, 2026

Keywords:

Basic sediment and water
Crude oil
Linear model
Machine learning
Prediction

Corresponding Author:

Jesi Pebralia,
Email: jesipebralia@unja.ac.id

ABSTRACT

Existing crude oil percentage prediction methods often rely on direct measurements and historical data, neglecting the coupled multiphase characteristics of oil–water–sediment systems, which limits predictive accuracy. This study develops a computational physics–based machine learning model integrating key multiphase production parameters, including water cut, basic sediment, and BS&W, using samples from PT. Pertamina Puspa Field Jambi. Data were split into two sets: one for model development and one for validation to prevent overfitting. Linear Regression, Support Vector Machine (SVM), and Random Forest algorithms were applied, with Linear Regression achieving the best performance. For the test dataset, the model yielded a Mean Absolute Error of 0.022168, a Mean Squared Error of 0.001227, and an accuracy of 0.99877, demonstrating precise capture of multiphase interactions. The proposed computational physics–based modelling framework provided improved predictive reliability and consistency. Correlation analyses indicated a coefficient of determination (R^2) of 0.99 and a perfect negative correlation ($r = -1$) between BS&W and oil content, showing that higher BS&W corresponds to lower oil percentage. This framework offers improved predictive reliability and consistency for crude oil quality assessment.

1. INTRODUCTION

Crude oil is a complex mixture containing thousands of different chemical compounds. The majority are hydrocarbons (molecules made up of hydrogen and carbon), but they also contain nitrogen, oxygen, sulfur, metals, and other elements in varying quantities (Saad et al., 2019). It is extracted from the earth using both traditional drilling methods and newer techniques like hydraulic fracturing. Once extracted, it's transported to refineries via pipelines, ships, or rail (Ramírez-Pradilla et al., 2019). In its raw state, crude oil isn't very useful to consumers. At refineries, it's separated into its individual components using a process called distillation. These components, or "fractions," can then be further processed to create a range of products (Ruble, 2019). Crude oil is the precursor to a vast array of products (Baruah & Tiwari, 2020). The most familiar are fuels like gasoline, diesel, and jet fuel (Wlazlowski, 2007). However, it's also the base for countless other products, including lubricating oils, asphalt, synthetic rubber, plastics, and certain chemicals (Okan et al., 2019). As a major source of energy, crude oil has significantly impacted economic development worldwide, enabling mobility and industrial growth (Fayomi et al., 2019; Litvinenko, 2020; Ngene et al., 2016; Onuoha & Elegbede, 2018; Qaim et al., 2020).

Oil and gas production requires continuous monitoring to accurately determine production quantities and quality parameters. Traditionally, the estimation of crude oil percentage in production streams has relied on geological and geophysical methods (Hanafy et al., 1997; Huffman, 2004; Sun et al., 2015), as well as direct laboratory measurements that require specialized equipment and operational time (Pebralia et al., 2022). However, these approaches are often limited in providing rapid predictive capability under varying production conditions. With advances in computational technology, machine learning (ML) has emerged as a promising alternative for predicting production parameters by learning complex relationships from historical operational data (Athambawa et al., 2023; Harlim et al., 2021; Mair et al., 2000; Masrom et al., 2023; Muneer et al., 2022; Seko et al., 2017). Unlike conventional measurement-based estimation, ML models enable indirect prediction of target variables using correlated production indicators, thereby improving prediction efficiency and operational decision-making (Alfian et al., 2023; Aman & Chhillar, 2023; Asadullah et al., 2023; Le et al., 2021).

Determining crude oil percentage in production streams remains a practical challenge, particularly in wells where complete measurement instrumentation is not always available. Several recent studies have demonstrated that machine learning (ML) techniques can effectively predict crude oil production characteristics using operational data, providing improved predictive efficiency compared to conventional estimation approaches (Chen et al., 2021; Obite et al., 2021; Okan et al., 2019; Raljević et al., 2021). However, many existing prediction models do not explicitly incorporate the multiphase flow dynamics of oil–water–sediment systems, which play a critical role in determining the final oil percentage during production. Motivated by this limitation, the present study proposes a computational physics–based machine learning framework that integrates key multiphase flow parameters, including water cut, basic sediment, and BS&W, to predict crude oil percentage. The proposed approach is applied to production data obtained from the Puspa structure of the PT Pertamina Jambi field to evaluate the effectiveness of multiphase flow–informed predictive modeling in improving oil percentage estimation. This work contributes by introducing a multiphase flow–informed machine learning modeling framework for crude oil percentage prediction, providing improved physical interpretability and predictive reliability compared to conventional data-driven approaches.

2. METHOD

2.1 Data Acquisition

This research employs primary data obtained from the Puspa structure of PT. Pertamina Jambi. The dataset was collected over a span of three months, from May to July 2023, and includes several parameters, specifically water cut, basic sediment, BS&W, and oil from crude oil. The data were sourced from five different wells, namely PPS-001, PPS-004, PPS-005-ST-ST, PPS-006, and PPS-007. Other materials used include water paste, demulsifier, and toluene, which are used to measure water cut and BS&W.

Two independent datasets were used to ensure objective model evaluation and to avoid overfitting. Dataset 1, as shown in Table 1, consisting of 83 samples, was employed for model development. After selecting the best-performing model based on evaluation metrics from dataset 1 treatment, a final validation step was conducted using dataset 2, as shown in Table 2, which contained 25 completely independent samples that were not involved in any stage of model training or preliminary testing. This two-stage evaluation strategy was implemented to ensure that the selected model demonstrates reliable generalization capability and can objectively predict unseen data.

The dataset 1 was divided into training and testing subsets using a train–test split procedure. In this study, 70% of the dataset was allocated for model training, while the remaining 30% was used for testing. The 70/30 partition was selected as a commonly adopted practice in supervised machine learning to provide a sufficient number of samples for model learning while maintaining an adequate independent dataset for unbiased performance evaluation (Géron, 2019; Hastie et al., 2008). In addition, this proportion is particularly suitable for small-to-medium-sized datasets, where preserving sufficient training samples is essential for model stability while ensuring representative testing evaluation.

Table 1 The top five and the last five of dataset 1

Source of Sample	Description	Data Collection Time	Data Analysis Time	WC (%)	BS (%)	BS&W (%)	Oil (%)
PPS-001	*Out Well*	18-Mei-23	18-Mei-23	97.97	0.03	98	2
PPS-001	*Out Well*	19-Mei-23	22-Mei-23	96.99	0.01	97	3
PPS-001	*Out Well*	20-Mei-23	22-Mei-23	97.99	0.01	98	2
PPS-001	*Out Well*	21-Mei-23	22-Mei-23	97.99	0.01	98	2
PPS-001	*Out Well*	23-Mei-23	24-Mei-23	79.99	0.01	80	20
…	...	…	…	…	…	…	…
PPS-007	*Out Well*	01-Jul-23	02-Jul-23	29.95	0.05	30	70
PPS-007	*Out Well*	04-Jul-23	05-Jul-23	47.95	0.05	48	52
PPS-007	*Out Well*	07-Jul-23	08-Jul-23	0.95	0.05	1	99
PPS-007	*Out Well*	16-Jul-23	24-Jul-23	22.99	0.01	23	77
PPS-007	*Out Well*	25-Jul-23	25-Jul-23	16.99	0.01	17	83

Table 2 The first five and the last five of dataset 2

Source of Sample	Description	Data Collection Time	Data Analysis Time	WC (%)
PPS-001	*Out Well*	17-Mei-23	17-Mei-23	96.95
PPS-001	*Out Well*	22-Mei-23	23-Mei-23	97.99
PPS-001	*Out Well*	01-Jun-23	01-Jun-23	89.99
PPS-001	*Out Well*	25-Jun-23	26-Jun-23	97.95
PPS-001	*Out Well*	19-Jul-23	24-Jul-23	96.95
…	...	…	…	…
PPS-007	*Out Well*	18-Mei-23	18-Mei-23	9.95
PPS-007	*Out Well*	04-Jun-23	04-Jun-23	17.99
PPS-007	*Out Well*	16-Jun-23	17-Jun-23	14.95
PPS-007	*Out Well*	19-Jul-23	24-Jul-23	29.99
PPS-007	*Out Well*	29-Jul-23	29-Jul-23	16.49

2.2 Research's Variables

The dependent variable in this study was the crude oil percentage in the gross production fluid, denoted as Y. The independent variables consist of the water cut (X_1), basic sediment (X_2), and basic sediment and water (X_3) contained in the gross fluid. These parameters were used as predictors to determine crude oil quality and percentage. The predictive relationship between the dependent and independent variables can be expressed in a general modeling form as follows in Equation (1):

$$Y = f(X_1, X_2, X_3) \tag{1}$$

where $f(.)$ represents the machine learning regression function used to estimate the crude oil percentage based on multiphase production parameters.

2.3 Data Analysis Technique

Data analysis was performed using a machine learning algorithm modeling, designed on the Jupyter Notebook platform using the Python language. The research process is illustrated in Figure 1.

The prediction used three modeling algorithms (Linear Model, SVM, and Random Forest). The performance analysis was then evaluated through model evaluation, and the best algorithm was selected to make predictions and classifications on the test data.

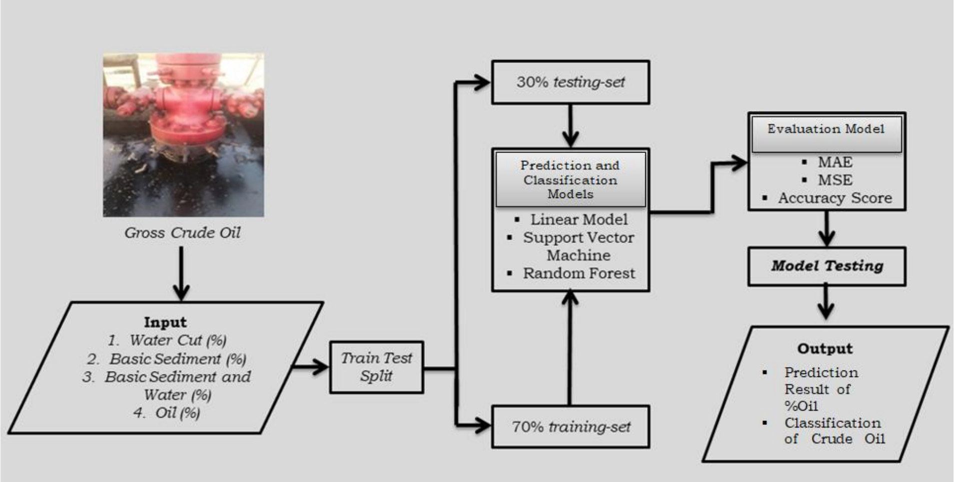

Figure 1 Framework of the Machine Learning Modeling Process

2.3.1 Linear regression model

In the linear regression approach, the relationship between multiphase flow parameters and crude oil percentage was approximated as a linear combination of input variables as expressed in Equation (2):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{2}$$

where β_0 is the intercept and β_j $(j = 1,2,3)$ are regression coefficients estimated by minimizing the Mean Squared Error (MSE). This model assumes a linear dependency between the oil percentage and the multiphase production indicators.

2.3.2 Support vector machine (SVM)

For nonlinear modeling, SVM seeks to determine a function of the form:

$$f(x) = w^T \varphi(x) + b \tag{3}$$

where $\varphi(x)$ is a nonlinear mapping function to a higher-dimensional feature space, w is the weight vector, and b is the bias term.

2.3.3 Random forest

Random forest (RF) is an ensemble learning method based on multiple decision trees. For T trees, the predicted crude oil percentage is obtained by averaging the outputs of individual trees:

$$y = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{4}$$

where $h_t(x)$ is the prediction from the t-th tree. Each tree is constructed using bootstrap sampling and random feature selection, allowing the model to capture complex nonlinear relationships between multiphase production parameters and oil percentage while reducing variance and overfitting.

2.4 Evaluation

The performance assessment of Machine Learning predictions was carried out by calculating the Mean Absolute Error (MAE), Mean Squared Error (MSE), and model accuracy values, as expressed in Equations (5), (6), and (7), respectively. MAE, MSE, and model accuracy are common evaluation methods used in machine learning to measure how well a model predicts the actual target or output values. The closer the MAE and MSE values are to 0, the better the model's performance in predicting data. Meanwhile, as the accuracy score of a machine learning model approaches, it indicates that the model is nearing perfect performance in classification or prediction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

$$Accuratation = 100\% - \%MAE \tag{7}$$

where n represents the number of data points, y_i is the actual value of the ith data, and $\hat{y}_i$ indicates the predicted value from the model for the ith data.

3. RESULTS AND DISCUSSION

3.1 Data Acquisition

The initial step was the analysis of data requirements, both for the reference data or training data and the data to be tested. Figure 2 provided a visual description of the relationship between the values of the measured variables (water cut, basic sediment, and water, oil) and the dataset numbers of the crude oil samples. These graphs assist in understanding the extent of variation or changes in the data variables relative to the dataset numbers representing the sequence of observations or records. In this context, it helps to observe how the levels of the measured variables (water cut, basic sediment, basic sediment and water, oil) change sequentially across the collected measurement data. This can provide insights into the magnitude of fluctuations in the percentages of the related parameters and the relationships among them.

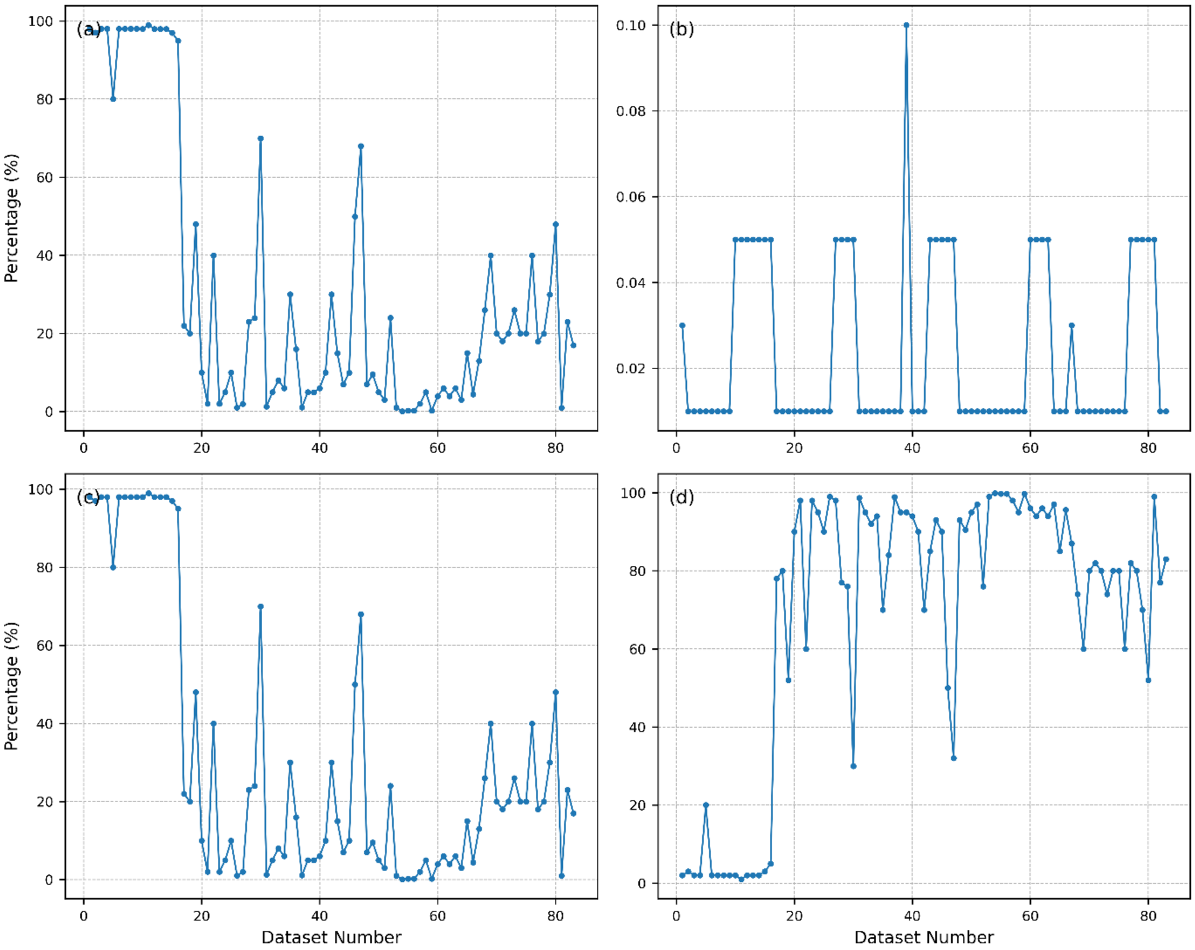

Figure 2 Distribution of multiphase production parameters across the dataset: (a) Water Cut, (b) Basic Sediment, (c) Basic Sediment and Water (BS&W), and (d) Oil percentage.

Table 3 Statistical description of the dataset

Parameters	Statistical Description			
	Min	Max	Mean	Std Dev
Water cut	0.09	98.95	31.28	35.02
Basic sediment	0.01	0.10	0.02	0.02
Water cut and Basic sediment (mix)	0.10	99.00	31.30	35.02
Oil (%)	1.00	99.90	68.70	35.02

From a multiphase flow perspective, as shown in Table 3, the statistical distribution of the variables reflects the conservation of phase fractions in the production fluid, with the total composition approximately equal to the sum of oil, water, and sediment fractions. The mean BS&W value (31.3%) closely complements the mean oil percentage (68.7%), indicating that the oil fraction is largely determined by the remaining portion of the multiphase mixture after accounting for water and sediment components. This consistency supports the physical interpretation that variations in water cut and sediment content directly control the observed fluctuations in crude oil percentage, reinforcing the relevance of multiphase flow dynamics as the underlying basis for the predictive modeling approach.

3.2 Linear Regression Test

The primary objective of regression analysis is to understand the cause-and-effect relationship between one or more independent variables (explanatory variables) and one dependent variable (the variable we wish to predict or explain). Figure 3 shows the results of the regression test on the training data.

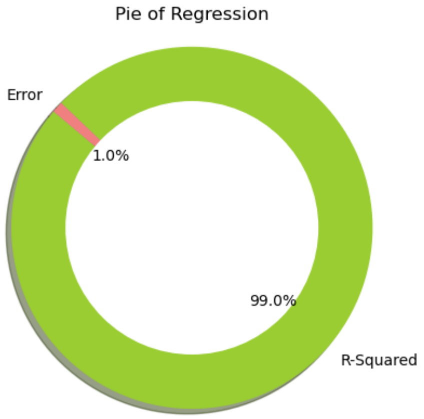

Figure 3 Regression test using Python.

Based on the regression results shown in Figure 3, the R² (coefficient of determination) value was 0.99. This indicates that the constructed regression model fits the data very well. In the context of regression analysis, an R^2 value of 0.99 suggests that nearly all the variation in the dependent variable (oil content) can be explained by variations in the independent variables (water cut, basic sediment, and BS&W). The resulting regression followed Equation (8).

$$y = 100 - 0{,}33X_1 - 0{,}33X_2 - 0{,}67X_3 \tag{8}$$

Equation (8) indicates that the oil percentage corresponds to the remaining fraction of the fluid mixture after accounting for the contributions of water and sediment phases. The negative coefficients for all independent variables show an inverse relationship, meaning that increases in water content, sediment content, or their combined fraction directly reduce the oil percentage in the produced fluid. The coefficients −0.33 for X_1 and X_2 suggest that a 1% increase in water cut or basic sediment individually decreases the oil percentage by approximately 0.33%, whereas the larger coefficient −0.67 for X_3 indicates that the combined water–sediment fraction has a more dominant influence on reducing the oil fraction. Overall, the structure of the equation is consistent with the conservation of multiphase fractions in production flow systems, where increases in non-oil components proportionally reduce the oil fraction in the mixture.

3.3 Data Training

At this stage, dataset 1 was used to train the model. These data served as examples to guide the model in learning patterns and generating predictions. The training data requires preprocessing before being used to train the model. Once the data variables to be trained have been adjusted, the next step is to perform a train-test split. Research conducted by Fashoto et al. (2021) indicates that using 70% of the data for training and 30% for testing yields the best empirical results. Therefore, this research divided the training data into 70% for the training set and 30% for the testing set. The training data were used to train the classification model, and the test data were used to evaluate the model's performance on data it had never "seen" before. The scikit-learn library was used to perform the train-test split.

3.4 Modelling

The modeling was conducted using three algorithms (Linear Model, Support Vector Machine, and Random Forest) utilizing the scikit-learn library. These algorithms were employed to predict and classify the training data. The modeling phase aimed to address the question, "How was the modeling constructed?" and to find solutions to the given problem. At this stage, the author trained each model to make predictions using the training data. The prediction results of the modeling on the training set are shown in Table 4.

Table 4 Prediction results of oil percentage from three different algorithms on the training set

Oil (%) (actual data)	LR Predict (%)	SVM Predict (%)	RF Predict (%)
98.7	98.7	98.60192	98.836
99.9	99.9	99.79973	99.433
85	85	84.94685	85.08
96	96	95.92679	95.984
98	98	97.90319	98.19
…	…	…	…
93	93	92.93226	93.34
94	94	93.93044	94.14
90	90	89.91778	90.15
94	94	93.93044	94.14
99.75	99.75	99.65	99.433

To test the performance of the modeling, error values and accuracy were calculated to assess the performance of the algorithm. In this case, the MAE (mean absolute error), MSE (mean square error), and model accuracy values were computed. Figure 4 shows a comparison of the values of the evaluation metrics for each model.

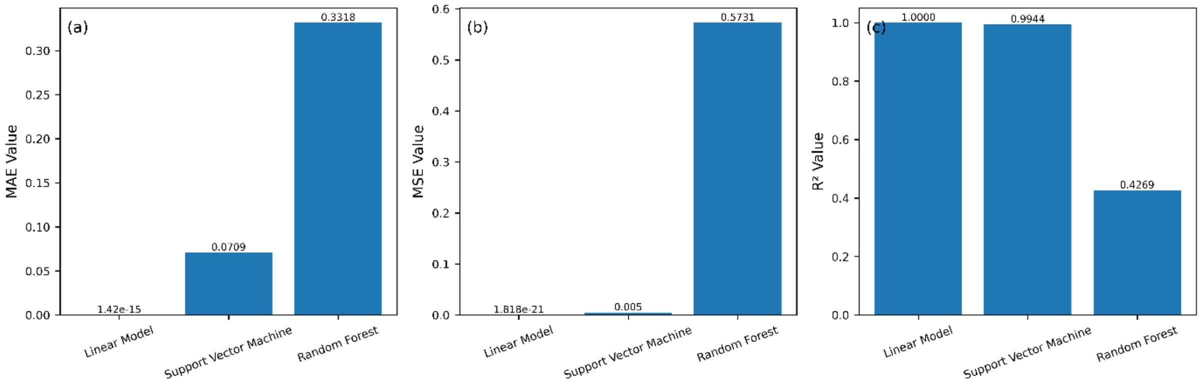

Figure 4 Comparative evaluation metrics of the three machine learning models: (a) Mean Absolute Error (MAE), (b) Mean Squared Error (MSE), and (c) Coefficient of Determination (R^2).

The smallest mean absolute error (MAE) value was held by the Linear Model, with a value of 1,421×10^{-15}, followed by the Support Vector Machine (SVM) with a value of 0.071, and lastly, the Random Forest with a value of 0.332, as shown in Table 5. Subsequent testing revealed that the Linear Model had the smallest MSE with a value of 1,818×10^{-21}, followed by the SVM model with a value of 0.005, and then the Random Forest model with a value of 0.573. Based on the modeling results, the best performance was achieved by the Linear Model, followed by SVM and Random Forest, with accuracy values of 1.000, 0.994, and 0.427, respectively. The Linear model had the smallest error value and the highest accuracy. Therefore, the Linear Model was chosen to predict the percentage of crude oil present in the test data.

Table 5 Algorithm Performance

Types of Algorithms	MAE	Evaluation MSE	Accuracy-Score
Linear Model	1.421×10^{-15}	1.818×10^{-21}	1.000
SVM	0.071	0.005	0.994
Random Forest	0.332	0.573	0.427

From Table 5, the results indicated that the Linear Model consistently outperformed the Support Vector Machine (SVM) and Random Forest models in predicting crude oil percentage, achieving near-zero MAE and MSE values and perfect accuracy. While nonlinear algorithms (SVM and Random Forest) are widely reported to perform well in complex and highly nonlinear systems, several studies have shown that linear regression models can yield superior or comparable performance when the underlying relationships among variables are governed by strong physical constraints (Abdalla et al., 2023; Fan et al., 2025).

From a physical perspective, this superior performance can be attributed to the deterministic relationships among the multiphase production parameters. Because BS&W captures nearly all variations in the non-oil fraction, the relationship between the independent variables and oil percentage is essentially linear. Consequently, the Linear Model is able to represent this relationship accurately with minimal error. In contrast, more complex nonlinear models like SVM or Random Forest are not required to capture these interactions and may introduce unnecessary variance when modeling a largely linear system.

Table 6 Data to be predicted for the oil percentage within it

No.	Water Cut
1	96.95
2	97.99
3	89.99
4	97.95
5	96.95
…	…
21	9.95
22	17.99
23	14.95
24	29.99
25	16.49

This outcome has important implications for oil-quality prediction. First, it demonstrates that incorporating physically meaningful multiphase parameters allows even simple linear models to achieve highly accurate predictions. Second, it highlights that understanding the underlying physics of fluid composition can guide model selection and prevent overcomplicating the predictive framework, a principle increasingly emphasized in physics-informed modeling approaches (Karniadakis et al., 2021). Finally, the strong inverse correlation between BS&W and oil percentage confirms that monitoring water cut and sediment content provides a reliable indirect measure of oil quality in field operations, consistent with established production-monitoring practices in petroleum engineering (Kamal et al., 2023).

3.5 Prediction and Classification of Data on Validation Stage

Following the comparative evaluation of the three algorithms, the Linear Model demonstrated the best predictive performance based on the evaluation metrics and was therefore selected for the validation stage. In this phase, the model was applied to an independent dataset to assess its generalization capability. Data preprocessing and variable transformation were performed to ensure compatibility with the model input structure, as summarized in Table 6.

After adjusting the variables in the test dataset, the next step was to predict the oil percentage for the samples contained in the test data. This stage aimed to predict the oil percentage in relation to its influence on the water cut percentage. In this case, there were two missing variables: basic sediment (BS) and basic sediment and water (BS&W). The algorithmic model constructed has been trained to predict the oil percentage value, even with some missing variables within it. Table 7 shows the predicted values of the crude oil percentage and the classification of oil quality based on company standards. The BS&W content set by the company for standard conditions is <0.5% of the total crude oil.

Table 7 Prediction results of the Linear Model algorithm on the Test Data

The source of samples	Description	Data Collection Time	Data Analysis Time	WC (%)	Oil (%)	Output
PPS-001	*Out Well*	17-Mei-23	17-Mei-23	96.95	3.017046	*Nonstandard*
PPS-001	*Out Well*	22-Mei-23	23-Mei-23	97.99	1.976834	*Nonstandard*
PPS-001	*Out Well*	01-Jun-23	01-Jun-23	89.99	9.978469	*Nonstandard*
PPS-001	*Out Well*	25-Jun-23	26-Jun-23	97.95	2.016842	*Nonstandard*
PPS-001	*Out Well*	19-Jul-23	24-Jul-23	96.95	3.017046	*Nonstandard*
…	...	…	…	…	…	...
PPS-007	*Out Well*	18-Mei-23	18-Mei-23	9.95	90.03483	*Nonstandard*
PPS-007	*Out Well*	04-Jun-23	04-Jun-23	17.99	81.99318	*Nonstandard*
PPS-007	*Out Well*	16-Jun-23	17-Jun-23	14.95	85.03381	*Nonstandard*
PPS-007	*Out Well*	19-Jul-23	24-Jul-23	29.99	69.99073	*Nonstandard*
PPS-007	*Out Well*	29-Jul-23	29-Jul-23	16.49	83.49349	*Nonstandard*

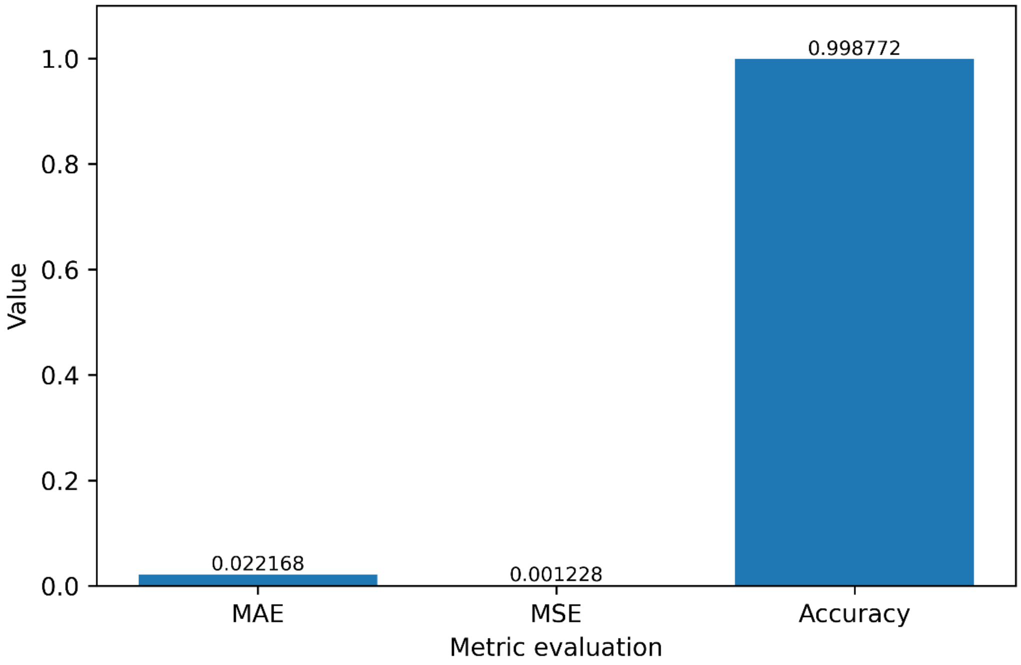

Figure 5 The metric performance of the predicted model (Linear model) on the validation data

The prediction and classification results yielded an MAE of 0.022168, MSE of 0.001228, and an accuracy value of 0.998772, as shown in Figure 5. These results indicated that the modelling constructed (Linear Model) performs well in producing accurate predictions. The MAE and MSE values indicate some level of relatively small errors and a trustworthy data accuracy level.

To further evaluate and analyze the strength of the relationships among all variables related to crude oil quality assessment, a correlation analysis was conducted. Correlation values, as shown in Table 8, assisted in measuring the strength and direction of the relationship between all variables. This provided an understanding of how changes in one variable relate to changes in another variable.

Table 8 Table of correlation relationships between variables

WC (%)	BS (%)	BS&W (%)	Oil (%)
1	0.18534	1	-1
0.18534	1	0.185894	-0.18589
1	0.185894	1	-1
-1	-0.18589	-1	1

Based on the correlation analysis presented in Table 8, the deterministic relationships observed among water cut (WC), BS&W, and oil percentage reflect the inherent mass-balance constraints in multiphase oil production systems. The perfect positive correlation between WC and BS&W, and the perfect negative correlation between these variables and oil percentage, therefore indicate definitional and compositional dependencies rather than purely statistical associations.

The correlation matrix indicates perfect correlations (±1) among some variables, notably between Water Cut (WC) and BS&W, as well as between these variables and Oil (%). Specifically, WC and BS&W are perfectly positively correlated (+1), while both show perfect negative correlation (−1) with Oil (%). Such strong correlations reflect a deterministic relationship rather than a purely statistical association. Physically, this is consistent with the conservation of multiphase fractions in the production fluid, where the sum of oil, water, and sediment fractions is approximately constant:

$$Oil + Water + Sediment \approx 100\% \tag{9}$$

Since BS&W represents the combined water and sediment fraction, the oil fraction can be expressed as follows, Equation (10):

$$Oil \approx 100 - BS + W \tag{10}$$

Moreover, given the relatively small variation in basic sediment, BS&W closely tracks water cut, explaining the perfect positive correlation between WC and BS&W, and the perfect negative correlation with oil (%). This demonstrates that the variables are physically linked through conservation principles.

A comparison with the results reported by Raljević et al. (2021) shows both methodological similarities and important differences in predictive performance. In their study, machine-learning multivariate regression models applied to NMR-derived compositional parameters achieved very high predictive accuracy (predicted $R^2 > 0.98$), demonstrating that crude-oil stability can be reliably modeled when rich physicochemical descriptors are available. Similar high-accuracy results have also been reported in spectroscopic and high-dimensional petroleum modeling studies, where machine learning algorithms such as Random Forest, Gradient Boosting, and Artificial Neural Networks were trained on compositional, rheological, or spectrometric features (Abdalla et al., 2023; Al-Mudhafar, 2020; Fan et al., 2025).

However, unlike those approaches that rely on complex laboratory-derived and high-dimensional feature spaces, the present study demonstrates that comparable predictive reliability can be achieved using simpler process-measurement variables (WC, BS, and BS&W) when strong physical constraints dominate the system behavior. This finding is consistent with studies emphasizing that model performance depends not only on algorithmic complexity but also on the structure of the underlying physical relationships (Carleo et al., 2019; Karniadakis et al., 2021). In systems governed by conservation principles and phase-balance equations, linear regression can outperform nonlinear models because it directly captures the dominant deterministic dependencies without introducing unnecessary model variance.

Furthermore, petroleum production forecasting studies have shown that machine learning models often benefit from integrating physically meaningful variables rather than relying solely on algorithmic sophistication (Al-Mudhafar, 2020). The results of the present study reinforce this perspective, when oil fraction is primarily controlled by the complementary relationship with water and sediment content, linear regression provides an efficient and interpretable solution with minimal computational burden. Therefore, while advanced spectroscopic-based multivariate modeling (Raljević et al., 2021) demonstrates the value of rich compositional inputs, the current findings highlight that field-accessible multiphase parameters can achieve similarly robust predictive performance when their physical interdependencies are properly leveraged.

4. CONCLUSION

This study demonstrated that the percentage of crude oil can be effectively predicted using multiphase composition parameters derived from Basic Sediment and Water (BS&W) measurements by comparing three machine-learning approaches (Linear Model, Support Vector Machine (SVM), and Random Forest). The results showed that the Linear Model provided the most reliable performance, with a mean absolute error (MAE) value of 0.022168, a mean squared error (MSE) of 0.001228, and an accuracy score of 0.99877. These findings indicated that the relationship between oil fraction and multiphase components was predominantly linear and physically consistent.

The correlation analysis further reinforces this conclusion. Given the relatively small variation in basic sediment within the dataset, BS&W closely tracks water cut, resulting in a perfect positive correlation between Water Cut (WC) and BS&W and a perfect negative correlation between these variables and Oil (%). This behavior confirmed that the variables were not independent but physically linked through multiphase conservation principles, where the total fluid composition must satisfy the approximate balance between oil and non-oil fractions. Consequently, the strong performance of the Linear Model reflected the deterministic nature of this conservation-based relationship rather than purely statistical coincidence.

The main contribution of this work lies in demonstrating that a simple, interpretable linear modeling framework, grounded in physical principles, can provide highly accurate predictions of crude oil quality. From a practical perspective, this approach enables rapid oil-percentage estimation, supports operational monitoring, and reduces reliance on time-consuming laboratory analysis.

However, several limitations should be acknowledged. The dataset size is relatively limited, and the strong physical dependencies among variables may restrict model generalization to production systems exhibiting more complex multiphase dynamics. Future research should therefore incorporate larger and more diverse datasets, include additional operational parameters such as temperature, pressure, and flow conditions, and explore hybrid physics-informed machine learning approaches to enhance robustness and predictive capability across varying reservoir environments.

ACKNOWLEDGEMENT

The author would like to thank PT Pertamina Hulu Rokan Zone 1 Field Jambi for granting permission and providing the data, making it possible to conduct this research effectively.

REFERENCES

Abdalla, R., Al-Hakimi, W., Perozo, N., & Jaeger, P. (2023). Real-Time Liquid Rate and Water Cut Prediction From the Electrical Submersible Pump Sensors Data Using Machine-Learning Algorithms. *ACS Omega*, *8*(14), 12671–12692. https://doi.org/10.1021/acsomega.2c07609

Al-Mudhafar, W. J. (2020). Integrating machine learning and data analytics for geostatistical characterization of clastic reservoirs. *Journal of Petroleum Science and Engineering*, *195*(April), 107837. https://doi.org/10.1016/j.petrol.2020.107837

Alfian, G., Saputra, Y. M., Subekti, L., Rahmawati, A. D., Atmaji, F. T. D., & Rhee, J. (2023). Utilizing deep neural network for web-based blood glucose level prediction system. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(3), 1829–1837. https://doi.org/10.11591/ijeecs.v30.i3.pp1829-1837

Aman, & Chhillar, R. S. (2023). Optimized stacking ensemble for early-stage diabetes mellitus prediction. *International Journal of Electrical and Computer Engineering*, *13*(6), 7048–7055. https://doi.org/10.11591/ijece.v13i6.pp7048-7055

Asadullah, M., Hossain, M. M., Rahaman, S., Amin, M. S., Sumy, M. S. A., Parh, M. Y. A., & Hossain, M. A. (2023). Evaluation of machine learning techniques for hypertension risk prediction based on medical data in Bangladesh. *Indonesian Journal of Electrical Engineering and Computer Science*, *31*(3), 1794–1802. https://doi.org/10.11591/ijeecs.v31.i3.pp1794-1802

Athambawa, A., Johar, M. G. M., & Khatibi, A. (2023). Behavioural intention to adopt cloud computing: A quantitative analysis with a mediatory factor using bootstrapping. *Indonesian Journal of Electrical Engineering and Computer Science*, *32*(1), 458–467. https://doi.org/10.11591/ijeecs.v32.i1.pp458-467

Baruah, B., & Tiwari, P. (2020). Effect of high pressure on nonisothermal pyrolysis kinetics of oil shale and product yield. *Energy and Fuels*, *34*(12), 15855–15869. https://doi.org/10.1021/acs.energyfuels.0c02538

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), 45002. https://doi.org/10.1103/RevModPhys.91.045002

Chen, H., Zhang, C., Jia, N., Duncan, I., Yang, S., & Yang, Y. Z. (2021). A machine learning model for predicting the minimum miscibility pressure of CO2 and crude oil system based on a support vector machine algorithm approach. *Fuel*, *290*, 1–27. https://doi.org/10.1016/j.fuel.2020.120048

Fan, D., Lai, S., Sun, H., Yang, Y., Yang, C., Fan, N., & Wang, M. (2025). Review of Machine Learning Methods for Steady State Capacity and Transient Production Forecasting in Oil and Gas Reservoir. *Energies*, *18*(4), 1–25. https://doi.org/10.3390/en18040842

Fayomi, O. S. I., Akande, I. G., & Odigie, S. (2019). Economic Impact of Corrosion in Oil Sectors and Prevention: An Overview. *Journal of Physics: Conference Series*, *1378*(2). https://doi.org/10.1088/1742-6596/1378/2/022037

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow* (2nd ed.). O'Reilly Media, Inc.

Hanafy, H. H., Macary, S. M., ElNady, Y. M., Bayomi, A. A., & El Batanony, M. H. (1997). A New Approach for Predicting the Cruide Oil Properties. *SPE Production Operations Symposium*, 439–452. http://www.onepetro.org/doi/10.2118/37439-MS

Harlim, J., Jiang, S. W., Liang, S., & Yang, H. (2021). Machine learning for prediction with missing dynamics. *Journal of Computational Physics*, *428*, 1–31. https://doi.org/10.1016/j.jcp.2020.109922

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Inc. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

Huffman, A. R. (2004). The future of pressure prediction using geophysical methods. *AAPG Memoir*, *76*, 217–233. https://doi.org/10.1306/m76870c19

Kamal, B., Abbasi, Z., & Hassanzadeh, H. (2023). Water-Cut Measurement Techniques in Oil Production and Processing—A Review. *Energies*, *16*(17). https://doi.org/10.3390/en16176410

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

Le, M. T., Vo, M. T., Pham, N. T., & Dao, S. V. T. (2021). Predicting heart failure using a wrapper-based feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, *21*(3), 1530–1539. https://doi.org/10.11591/ijeecs.v21.i3.pp1530-1539

Litvinenko, V. S. (2020). Digital Economy as a Factor in the Technological Development of the Mineral Sector. *Natural Resources Research*, *29*(3), 1521–1541. https://doi.org/10.1007/s11053-019-09568-4

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., & Webster, S. (2000). An Investigation of machine learning based prediction systems. *Journal of Systems and Software*, *53*(1), 23–29. https://doi.org/10.1016/S0164-1212(00)00005-4

Masrom, S., Rahman, R. A., Baharun, N., Rohani, S. R. S., & Rahman, A. S. A. (2023). Machine learning with task-technology fit theory factors for predicting students' adoption in video-based learning. *Bulletin of Electrical Engineering and Informatics*, *12*(3), 1666–1673. https://doi.org/10.11591/eei.v12i3.5037

Muneer, A., Ali, R. F., Alghamdi, A., Taib, S. M., Almaghthawi, A., & Abdullah Ghaleb, E. A. (2022). Predicting customers churning in banking industry: A machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 539–549. https://doi.org/10.11591/ijeecs.v26.i1.pp539-549

Ngene, S., Tota-Maharaj, K., Eke, P., & Hills, C. (2016). To cite this article: Stanley Ngene, Kiran Tota-Maharaj, Paul Eke, Colin Hills. Environmental and Economic Impacts of Crude Oil and Natural Gas Production in Developing Countries. *International Journal of Economy, Energy and Environment*, *1*(3), 64–73. https://doi.org/10.11648/j.ijeee.20160103.13

Obite, C. P., Chukwu, A., Bartholomew, D. C., Nwosu, U. I., & Esiaba, G. E. (2021). Classical and machine learning modeling of crude oil production in Nigeria: Identification of an eminent model for application. *Energy Reports*, *7*, 3497–3505. https://doi.org/10.1016/j.egyr.2021.06.005

Okan, M., Aydin, H. M., & Barsbay, M. (2019). Current approaches to waste polymer utilization and minimization: a review. *Journal of Chemical Technology and Biotechnology*, *94*(1), 8–21. https://doi.org/10.1002/jctb.5778

Onuoha, M. E., & Elegbede, I. O. (2018). The Oil Boom Era: Socio-Political and Economic Consequences. In *The Political Ecology of Oil and Gas Activities in the Nigerian Aquatic Ecosystem*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-809399-3.00006-9

Pebralia, J., Amri, I., & Rifa'i, A. I. (2022). Measuring convective heat transfer in a room equipped with an air conditioner. *Physics Education*, *57*(5). https://doi.org/10.1088/1361-6552/ac832e

Qaim, M., Sibhatu, K. T., Siregar, H., & Grass, I. (2020). Environmental, economic, and social consequences of the oil palm boom. *Annual Review of Resource Economics*, *12*, 321–344. https://doi.org/10.1146/annurev-resource-110119-024922

Raljević, D., Parlov Vuković, J., Smrečki, V., Marinić Pajc, L., Novak, P., Hrenar, T., Jednačak, T., Konjević, L., Pinević, B., & Gašparac, T. (2021). Machine learning approach for predicting crude oil stability based on NMR spectroscopy. *Fuel*, *305*(August). https://doi.org/10.1016/j.fuel.2021.121561

Ramírez-Pradilla, J. S., Blanco-Tirado, C., Hubert-Roux, M., Giusti, P., Afonso, C., & Combariza, M. Y. (2019). Comprehensive Petroporphyrin Identification in Crude Oils Using Highly Selective Electron Transfer Reactions in MALDI-FTICR-MS. *Energy and Fuels*, *33*(5), 3899–3907. https://doi.org/10.1021/acs.energyfuels.8b04325

Ruble, I. (2019). The U.S. crude oil refining industry: Recent developments, upcoming challenges and prospects for exports. *Journal of Economic Asymmetries*, *20*(August 2019), e00132. https://doi.org/10.1016/j.jeca.2019.e00132

Saad, M. A., Kamil, M., Abdurahman, N. H., Yunus, R. M., & Awad, O. I. (2019). An overview of recent advances in state-of-the-art techniques in the demulsification of crude oil emulsions. *Processes*, *7*(7). https://doi.org/10.3390/pr7070470

Seko, A., Hayashi, H., Nakayama, K., Takahashi, A., & Tanaka, I. (2017). Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, *95*(14), 1–11. https://doi.org/10.1103/PhysRevB.95.144110

Sun, L., Fang, C., Li, F., Zhu, R., Zhang, Y., Yuan, X., Jia, A., Gao, X., & Su, L. (2015). Innovations and challenges of sedimentology in oil and gas exploration and development. *Petroleum Exploration and Development*, *42*(2), 143–151. https://doi.org/10.1016/S1876-3804(15)30001-X

Wlazlowski, S. (2007). Crude oil — end-product linkages in the European petroleum markets. *OPEC Review*, *31*(2), 73–90. https://doi.org/10.1111/j.1468-0076.2007.00177.x