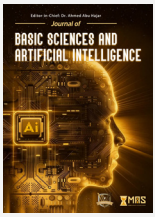
Journal of Basic Sciences and Artificial Intelligence (JBSAI)

DOI: 10.65098/jbsai.01.2026.08.17



RESEARCH ARTICLE

Beyond Prediction Accuracy: Evaluating Machine Learning Generalization for Crude-Oil Wax Precipitation

Yaman Hamed*, Eng Hao Louis Tan

Department of Applied Science, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia
Corresponding Author E-mail: yaman.hamed@utp.edu.my

ARTICLE DETAILS

Article History:
Received 02 May 2026
Accepted 20 Aug 2026
Available online 31 Aug 2026

Online Article Code

ABSTRACT

Wax precipitation alters crude-oil flow behaviour and provides the solid material that may later form pipeline deposits. Machine-learning models can accurately predict wax weight, but their ability to generalize beyond the training domain remains uncertain. This study evaluated artificial neural network (ANN), decision tree, random forest, and XGBoost models using 550 PVTSim-generated cases representing five crude-oil compositions. Model inputs included temperature, pressure, crude class, and the carbon fractions C1–C6, C7–C20, and C21+. Two validation strategies were employed: prediction under unseen pressure conditions within the represented crude compositions and leave-one-crude-out validation to assess transferability to unseen crude compositions. All models achieved R^2 above 0.999 for unseen pressure conditions, with ANN producing the lowest root mean square error (RMSE) of 0.0361 weight percent (wt.%) and a normalized RMSE of 0.261%. Performance declined substantially when an entire crude composition was excluded from training, and no model generalized reliably across all five crude compositions. SHapley Additive exPlanations analysis identified temperature as the dominant predictor. Heavy crude oils, characterized by lower light-end (C1–C6) and higher heavy-end (C21+) fractions, consistently produced higher predicted wax weight than light crude oils, whereas pressure had only a weak influence within the simulated range.

KEYWORDS

Wax Precipitation, Wax Weight, Crude Oil Composition, Machine Learning, Cross-Crude Validation, SHapley Additive exPlanations

1. INTRODUCTION

Wax precipitation is a persistent flow-assurance concern during crude-oil production and transport. Paraffinic components become less soluble as the fluid cools. Solid crystals appear when the temperature reaches the wax appearance temperature. The solid fraction then increases as cooling continues. The precipitated wax may increase viscosity, promote gel formation, and provide material for deposition on cooled surfaces (Aiyejina et al., 2011; Alnaimat & Ziauddin, 2020; Kiyingi et al., 2022).

Wax precipitation and wax deposition are related but distinct. Precipitation is a phase-equilibrium process. Deposition also requires transport toward the wall, adhesion, observation, and ageing of the deposited layer. Wall temperature, thermal gradients, flow rate, shear, pipe geometry, and deposit composition affect the final deposit thickness and strength (Singh et al., 2000; Huang et al., 2011; Li et al., 2024). The target in this study is simulated wax weight percentage. It represents the equilibrium amount of wax formed under a given pressure, temperature, and crude composition. However, it is not a direct measure of wall-deposit thickness or deposition rate.

Crude composition strongly controls the amount of wax formed. Long-chain normal paraffins are the main wax-forming species, so crude density or broad light-heavy classification alone does not adequately describe wax tendency. Heavy oils, for example, may contain large resin or asphaltene fractions without having the same paraffinic distribution. Thus, Carbon-number fractions therefore provide more useful compositional information than a simple light-heavy label. The temperature (ambient and/or inline) controls the immediate precipitation response, while the pipeline pressure can alter phase behaviour through changes in solubility and fluid composition. However, the magnitude of the pressure effect remains system-dependent (Lira-Galeana et al., 1996; Pan et al., 1997).

When laboratory measurements are limited, thermodynamic simulators can generate wax- appearance and wax-fraction responses across broad operating ranges. A data-driven surrogate can then reproduce these simulated outputs at lower computational cost, making it useful for screening, optimization, and repeated preventive-planning calculations. However, the findings are affected by the limitations of simulated data because the surrogate remains bounded by the simulator assumptions and by the crude compositions. Therefore, its real-field validity remains limited without measured field or laboratory data. Nevertheless, within the represented simulation domain, the surrogate remains useful for early screening and for supporting preventive planning ahead of potential wax-precipitation problems.

Machine-learning studies have predicted wax-deposition potential, precipitated wax amount, deposit thickness, deposition rate, and deposit location. Early studies relied mainly on artificial neural networks. Later work introduced support-vector methods, tree ensembles, hybrid models, and explainable machine learning (Kamari et al., 2013; Amar et al., 2022; Amiri-Ramsheh et al., 2023; Ahmadi, 2025).

The central problem is not whether machine-learning models can fit a wax-weight dataset, but whether they remain accurate when applied to an operating condition or crude composition absent from training. A point-wise random split can place neighbouring temperatures from the same pressure curve, or samples from the same crude composition, in both the training and testing sets. Such evaluation mainly measures interpolation within known response surfaces rather than transfer to a new crude assay.

Model interpretation presents a second challenge. A high feature-importance value identifies influential variables but does not indicate whether they increase or decrease wax weight. This is particularly important for correlated composition descriptors. In the present dataset, C1–C6, C7– C20, C21+, and crude class vary together across the five crude compositions. Their importance therefore must be interpreted together with the direction of their effects and the physical limitations of the dataset.

The reviewed literature shows continued progress in prediction accuracy, uncertainty analysis, and explainability. Ahmadi (2025) compared random forest (RF), XGBoost, and physics-informed neural networks using SHapley Additive exPlanations (SHAP) and bootstrap analysis on compositional data. Sarkodie et al. (2026) compared five algorithms for deposition-rate prediction and converted the best-performing black- box model into an interpretable surrogate. However, neither study reported a validation design that excluded an entire crude composition from model development. The transferability of high- performing wax-prediction models across crude types therefore remains insufficiently tested.

Based on the foregoing discussion, this study addresses the gap in cross-crude validation through three objectives: (1) to develop and compare ANN, decision tree (DT), RF, and XGBoost models for predicting simulated wax weight; (2) to evaluate prediction under two validation conditions, complete pressure curves excluded from training and complete crude compositions excluded from training; and (3) to quantify the magnitude and direction of operating and compositional effects using SHAP analysis. The main contribution is a dual validation framework that separates in-domain operating-condition interpolation from cross-crude composition transfer.

2. LITERATURE REVIEW

2.1 Physics of Wax Weight and Factor Effects

Thermodynamic wax models describe solid-liquid equilibrium in multicomponent hydrocarbon mixtures. Their performance depends on heavy-end characterisation, the solid-phase model, and the representation of paraffin distributions. Lira-Galeana et al. (1996) developed a molecular thermodynamic framework for wax precipitation in petroleum mixtures. Pan et al. (1997) examined pressure and composition effects and showed that precipitated wax depends on heavy n-paraffin concentration. These studies established the direct role of carbon-number distribution in wax formation.

Temperature is known to be the dominant operating variable because cooling reduces the solubility of higher molecular-weight paraffins. An important question is whether the data-driven models reproduce this established physical behaviour. Pressure effects are also known to depend on the fluid system and phase region. Although pressure can alter the phase boundary and the distribution of lighter components, its influence may remain small over a restricted liquid-phase range. Reviews by Aiyejina et al. (2011), Alnaimat and Ziauddin (2020), Kiyingi et al. (2022), and Li et al. (2024) show that composition, temperature, pressure, flow, and wall conditions act at different stages of precipitation and deposition. The SHAP analysis presented in this study therefore provides a means of assessing whether the trained machine-learning models capture these established physical relationships rather than relying solely on statistical correlations.

Although closely related, total precipitated wax and wall deposition represent different prediction targets. Bulk wax weight is governed primarily by phase equilibrium, whereas deposit thickness and deposition rate are also influenced by transport, wall heat transfer, hydrodynamics, adhesion, shear removal, and ageing. Consequently, a model developed to predict wax precipitation weight should not be interpreted as a direct predictor of pipeline blockage or pigging interval. The term wax precipitation weight is therefore used throughout this study to distinguish the prediction target from downstream deposition behaviour.

2.2 Machine-Learning Modelling of Wax-Related Quantities

Obanijesu and Omidiora (2008) developed an ANN for wax-deposition potential using reservoir pressure-volume-temperature and viscosity information. Kamari et al. (2013) applied a coupled simulated-annealing least-squares support-vector machine to estimate wax amount and included an applicability-domain check. Behbahani et al. (2015) used a multilayer perceptron (MLP) to predict wax appearance temperature and precipitated wax, then compared the predictions with thermodynamic models. Collectively, these studies demonstrated that nonlinear data-driven methods can reproduce wax-related quantities from relatively small petroleum datasets.

Subsequent studies shifted their focus from bulk wax precipitation to pipeline deposition. Jalalnezhad and Kamali (2016) predicted deposit thickness in turbulent flow. Xie and Xing (2017) used an Radial Basis Function (RBF) neural network for deposition-rate prediction based on thermal and flow variables. Yao et al. (2021) developed a time-dependent data-mining approach for pipeline deposits. Kim et al. (2022) used simulation data and a stacked autoencoder to estimate the location and amount of wax in a pipeline. These targets require flow and pipeline descriptors that are not available in the present dataset.

Among the published studies, wax-weight prediction is most closely aligned with the present work. Amar et al. (2022) used 88 experimental measurements and two MLP training schemes. Inputs included temperature, specific gravity, and six carbon-number fractions. The best model achieved an root mean square error (RMSE) of 0.2198 and an R^2 of 0.9971. Amiri-Ramsheh et al. (2023) compared MLP, RBF, Cascade-Forward Neural Network (CFNN), and Generalized Regression Neural Network (GRNN) models using 173 literature measurements with American Petroleum Institute (API) gravity and pour point as inputs. These studies demonstrated that nonlinear models can accurately estimate wax amount, but the comparisons were limited to different neural-network architectures rather than broader model families that include tree-based methods.

More recent studies have emphasized model interpretability. Ahmadi (2025) compared RF, XGBoost, and physics-informed neural networks using 88 compositional samples, with SHAP and partial dependence analysis used to interpret the effects of temperature and hydrocarbon fractions. Sarkodie et al. (2026) used 215 experimental and field observations to predict deposition rate, comparing five black-box models before developing an elastic-net surrogate for interpretation. Although these studies improved explainability, both relied on conventional train-test or cross-validation procedures, and neither evaluated model performance using complete crude- composition holdout.

Table 1 summarises the main studies that position the present work within the existing literature. Three observations emerge from the reviewed studies. Machine learning can predict wax-related quantities. The target definition varies substantially across studies. Validation is usually performed within the available sample pool, which does not directly test transfer to a crude composition absent from training.

2.3 Critical Synthesis and Research Gap

The first limitation concerns the validation structure. Wax datasets often contain repeated measurements from the same crude or smooth points from the same simulated response curve. Random splitting assumes that observations are independent, an assumption that

Table 1 Comparison of Selected Machine-Learning Studies on Wax Prediction

Study	Target and data	Model (s)	Validation and interpretation	Main limitation
Obanijesu and Omidiora (2008)	Wax-deposition potential; 11 Nigerian reservoirs	ANN	Within-dataset prediction	Small field-specific set; no cross-crude holdout
Kamari et al (2013)	Wax amount; experimental literature data	CSA-LSSVM	Train-test modelling and applicability-domain check	No whole-crude transfer test
Behbahani et al (2015)	WAT and precipitated wax; experimental and literature data	MLP and thermodynamic models	Accuracy comparison with physical models	No structured crude-level validation
Jalalnezhad and Kamali (2016)	Deposit thickness; turbulent-flow data	Neuro-fuzzy model	Evaluation within reported operating range	Single model; limited transfer evidence
Xie and Xing (2017)	Deposition rate; laboratory-derived factors	RBF neural network	Within-dataset error analysis	Composition not represented
Yao et al (2021)	Time-dependent pipeline deposits; operating data	Time-dependent data mining	Time-dependent evaluation	Pipeline-specific; not a wax-weight model
Kim et al (2022)	Deposit location and amount; simulation data	Stacked autoencoder	Simulation-based train-test evaluation	No transfer to a different crude composition
Amar et al (2022)	Wax weight; 88 experimental measurements	MLP-LMA and MLP-BR	Conventional data partition; high R^2	One model family; no whole-crude holdout
Amiri-Ramsheh et al (2023)	Total wax content; 173 literature values	MLP, RBF, CFNN, GRNN	Random train-test split	Only neural models; two broad crude descriptors
Ahmadi (2025)	Wax percentage; 88 compositional samples	RF, XGBoost, PINN	Bootstrap uncertainty, SHAP, partial dependence	No complete crude-composition holdout reported
Sarkodie et al (2026)	Deposition rate; 215 experimental and field observations	SVR, RF, MLP, GB, KNN and surrogate	80/20 split, five-fold Cross-Validation (CV), interpretable surrogate	No fluid-group holdout; different deposition-rate target

becomes weak when several observations share one fixed composition and differ only in temperature or pressure. Structured cross-validation is therefore required when observations form groups or response surfaces, as random splitting can produce optimistic estimates of model transferability (Roberts et al., 2017).

The second limitation concerns the distinction between interpolation and extrapolation. A model may accurately estimate an unobserved pressure for a known crude because the composition remains within its training domain. Predicting a new crude is more challenging because the model must transfer to a composition that has never been observed. Existing wax studies rarely distinguish between these two tasks. Consequently, the reported R^2 values do not indicate whether a model can generalize across crude types.

The third limitation relates to interpretation under correlated features. Recent SHAP-based studies have improved the interpretation of temperature and carbon-fraction effects, representing a substantial advance over impurity-based importance measures. However, explainability does not remove the need for rigorous validation. A physically plausible explanation cannot compensate for poor transfer to an unseen crude, just as high predictive accuracy cannot compensate for implausible factor relationships.

Taken together, these limitations define a clear research gap. The reviewed literature does not combine ANN, DT, RF, and XGBoost within a dual validation framework consisting of complete operating-condition holdout and complete crude-composition holdout. It also does not integrate this validation strategy with directional SHAP analysis for simulated wax weight. The present study addresses this gap by separating operating-condition interpolation from cross-crude composition transfer while relating both to physically interpretable factor effects.

3. MATERIALS AND METHODS

3.1 Data Source and Variables

The data were generated using PVTSim 20 through its wax P-T simulation framework. The source dataset contained five crude-oil compositions labelled Tapis, crude 1, crude 3, crude 4, and crude 5. The absence of a case labelled crude 2 was retained from the source data because the available records do not establish whether Tapis replaced crude 2 or was defined independently. For each crude composition, wax formation was simulated at 1 bar and from 5 to 50 bar at 5-bar increments, with 10 temperature-dependent wax-weight outputs at each pressure. This produced 110 observations per crude composition and 550 observations in total.

The model inputs were temperature (°C), pressure (bar), crude class, C1–C6 fraction (%), C7–C20 fraction (%), and C21+ fraction (%), with wax weight percentage as the output. Crude class was coded as 0 for the three lighter crude oils and 1 for the two heavier crude oils. The carbon fractions remained fixed within each crude composition, while temperature and pressure varied systematically.

The five crude compositions covered a broad compositional range, with C1–C6 fractions ranging from 9.720% to 77.065% and C21+ fractions from 0% to 32.172%. The maximum simulated wax weight was 13.865%. Crude 4 and crude 5 had the highest C21+ fractions and produced the highest wax-weight values, whereas Tapis and crude 3 had lower C21+ fractions and correspondingly lower wax weights. Because the compositional variables were fixed within each crude composition, crude class and the carbon fractions were strongly correlated, and this dependency was considered during model interpretation (Figure 1).

3.2 Validation Design

The dataset comprises complete wax-response curves for each crude-pressure combination. An observation-wise random split would place

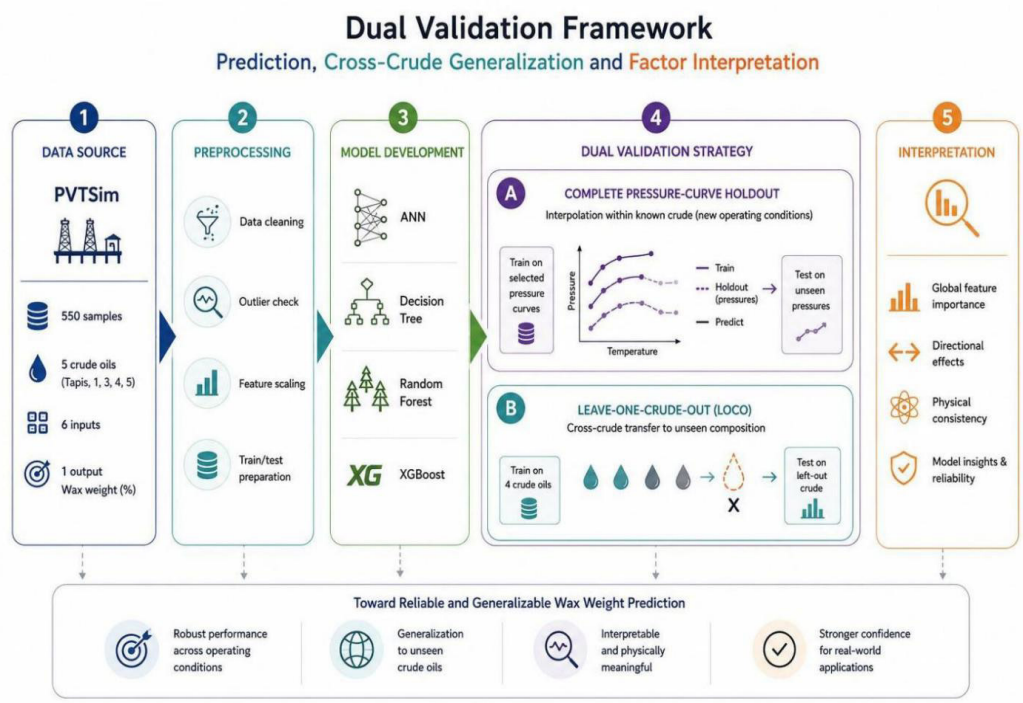


Figure 1 Dual Validation Study Framework

neighbouring temperatures from the same curve in both the training and testing sets while retaining every crude composition in both subsets. Two structured validation pathways were therefore adopted.

The first pathway assessed prediction at unseen pressure conditions within represented crude compositions. All observations at 5, 25, and 45 bar were excluded from model development, producing 150 test observations comprising three complete pressure curves for each of the five crude oils. The remaining 400 observations formed the training pool. Grouped five-fold cross-validation was applied within this pool, with each crude-pressure curve treated as one group so that all observations from the same curve remained in a single fold. Hyperparameters were selected using grouped cross-validation RMSE.

The pressure-held-out test evaluates interpolation across pressure. The test pressures were absent from training, whereas each crude composition remained represented. This provides a relevant surrogate test for repeated simulation within a known crude domain but does not evaluate prediction for a new crude assay.

The second pathway used leave-one-crude-out (LOCO) validation. One complete crude composition, including all 110 temperature-pressure observations, was excluded from model development. The model was then refitted using the remaining four crude oils and evaluated on the excluded composition. This procedure was repeated for Tapis and crude oils 1, 3, 4, and 5. The hyperparameters identified in the first pathway were fixed throughout the LOCO refits to maintain identical model configurations across the five crude-composition tests. Because these hyperparameters were selected using a development pool containing all five crude compositions, the LOCO analysis represents cross-crude transfer under fixed globally selected configurations rather than fully nested cross-crude model selection.

LOCO validation evaluates transfer across crude composition. Temperature and pressure remained within the ranges represented during training, whereas the combination of C1–C6, C7–C20, C21+, and crude class was absent from the training data. Only five crude-level tests were available; therefore, the results were analysed descriptively. An inferential model- comparison test was not applied because five heterogeneous crude domains do not provide sufficient independent units for a reliable rank-based comparison.

3.3 Machine-Learning Models and Tuning

Four nonlinear supervised regression models were evaluated: ANN, DT, RF, and XGBoost. ANN represented a flexible nonlinear connectionist model. DT provided a single rule-based tree. RF represented a bagged tree ensemble (Breiman, 2001), while XGBoost represented regularised gradient boosting (Chen & Guestrin, 2016). Multiple linear regression (MLR) was included as a simple baseline to assess whether the predictive accuracy of the nonlinear models could be attributed primarily to the smoothness of the response surface.

Moderate hyperparameter tuning was performed to obtain stable models without an excessive search that would add limited value to a smooth simulation dataset. Randomized search was used to minimise grouped cross-validation RMSE. ANN inputs were standardised within the modelling pipeline, whereas tree-based models used the original feature scales. Early stopping was applied to ANN training. The selected hyperparameter configurations are listed in Table 2.

The Randomized search used 20 iterations for DT and 24 iterations for ANN, RF, and XGBoost, with a fixed random seed of 42. Search ranges were: ANN, 2–3 hidden layers, 8–32 batch size, and 0.001–0.005 learning rate; DT, depth 6–14; RF, 100–400 trees and depth 6–12; and XGBoost, 150–300 trees, depth 3–6, and learning rate 0.03–0.10. Modelling was implemented in Python 3.13.5 using scikit-learn 1.8.0 and XGBoost 3.1.3.

The same feature set was used for all models to ensure a fair comparison.

Table 2 Model Configurations after Grouped Cross-Validation

Model	Selected configuration	Search configuration
ANN	Three hidden layers: 100–50–25 neurons; Rectified Linear Unit (ReLU); Adam; learning rate 0.003; batch size 16; alpha 0.001; early stopping	2–3 hidden layers; batch size 8–32; learning rate 0.001–0.005; 24 iterations
DT	Maximum depth 12; minimum leaf size 2; minimum split size 2; all features considered at each split	Depth 6–14; 20 iterations
RF	200 trees; maximum depth 8; minimum leaf size 2; minimum split size 2; all features considered	100–400 trees; depth 6–12; 24 iterations
XGBoost	220 trees; depth 5; learning rate 0.08; subsample 1.0; column sample 0.8; min child weight 3; alpha 0.01; lambda 5	150–300 trees; depth 3–6; learning rate 0.03–0.10; 24 iterations

Reproducibility settings: Random seed = 42; Python 3.13.5; scikit-learn 1.8.0; XGBoost 3.1.3.

No feature was removed based on importance calculated from the full dataset because this would introduce information from the test domain into model development. Retaining a fixed feature set also allowed the cross-crude validation to reflect the true consequence of extrapolating to unseen crude compositions.

3.4 Performance Measures and Statistical Comparison

Model performance was assessed using the coefficient of determination (R^2), RMSE, mean absolute error (MAE), and range-normalized RMSE (NRMSE). RMSE and MAE retain the unit of wax weight percentage, whereas NRMSE expresses RMSE as a percentage of the observed target range in the corresponding test set. For LOCO validation, the denominator was the wax-weight range of the held-out crude. MAPE was not used because zero wax values make percentage errors unstable. The uncertainty in pressure-held-out performance was quantified using 95% bootstrap confidence intervals for RMSE, with complete crude-pressure curves resampled as groups to preserve the dependence structure within each curve.

For the pressure-held-out pathway, curve-level MAE was calculated for each of the 15 excluded crude-pressure curves. Friedman (1937) test was used to compare the four models across these curves. Pairwise Wilcoxon signed-rank tests with Holm adjustment were performed only if the overall Friedman test was significant. MAE was selected for the statistical comparison because it provides a robust curve-level summary that is less sensitive to isolated large errors, whereas RMSE was retained as the optimisation criterion during hyperparameter tuning. The LOCO pathway was summarised using per-crude RMSE, MAE, NRMSE, and R^2. No inferential comparison was applied because only five crude-level test cases were available.

3.5 Feature Interpretation

RF was selected for SHAP interpretation because its pressure-held-out accuracy was close to ANN and its tree structure supports exact TreeSHAP calculations. Mean absolute SHAP values were used for global importance. Directional dependence plots were generated for temperature, C1–C6, and C21+ (Lundberg & Lee, 2017).

SHAP values describe model contributions, not causal effects. This distinction is important because crude class and the three carbon fractions are strongly correlated and take only five composition levels. A feature can receive a high SHAP magnitude because it identifies a crude group. The direction plots were therefore interpreted as model-consistent associations and checked against established wax physics. They were not treated as independent experimental effects (Janzing et al., 2020).

SHAP analysis used the RF model evaluated on the pressure-held-out test set. It was not applied to the LOCO predictions. Explaining unstable extrapolation to an unseen composition could give a precise-looking interpretation for a model operating outside its reliable domain.

4. RESULTS AND DISCUSSION

4.1 Wax-Weight Response Across Crude Compositions

The simulated curves showed the expected increase in wax weight during cooling. Wax weight was zero or near zero at the highest temperatures and increased as temperature declined below the wax precipitation region. The response was smooth for every crude-pressure combination, explaining why nonlinear models can achieve very small in-domain prediction errors.

The magnitude of wax precipitation differed substantially among the crude oils. Crude 4 reached the highest wax weight, followed by crude 5. These two crude oils contained the highest C21+ fractions and the lowest C1–C6 fractions. In contrast, Tapis and crude 3 produced much lower wax weights, whereas crude 1 occupied an intermediate range. This pattern indicates that temperature governs movement along each curve, whereas crude composition determines the overall level and shape of the response.

Pressure influenced the simulated wax weight within each crude composition, although its effect was generally smaller than the differences among crude oils. The pressure curves remained relatively close compared with the large compositional differences, indicating that crude composition dominated the overall response in this dataset. This observation should not be generalised beyond the simulated pressure range or to other fluid systems. Figure 2 presents temperature in descending order to reflect the physical cooling process, making the progressive increase in wax weight easier to interpret. For visualization purposes, the pressures 1, 25, and 50 bar were selected as representative low-, intermediate-, and high-pressure conditions and should not be confused with the 5, 25, and 45 bar levels used for pressure-held-out validation.

4.2 Prediction at Unseen Pressure Conditions

Table 3 summarizes grouped cross-validation and pressure-held-out performance. The four nonlinear models achieved R^2 above 0.999 on the excluded pressure curves, whereas the MLR baseline achieved

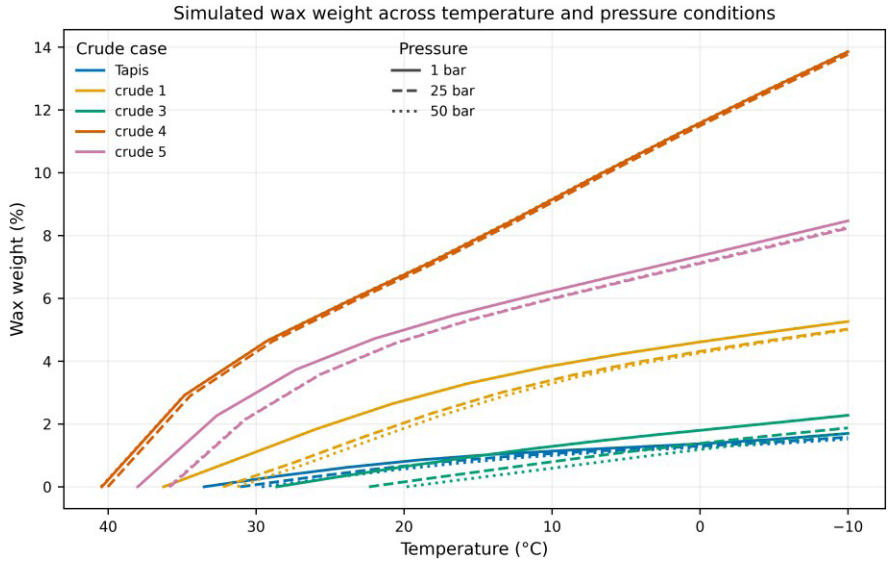


Figure 2 Wax-Weight Response to Cooling at Selected Pressures

Table 3 Model Performance on Complete Pressure Curves Excluded from Model Development

Model	CV RMSE	CV SD	Test RMSE	95%CI	Test MAE	NRMSE (%)	Test R^2
MLR (baseline)	1.2017	0.1098	1.1957	0.9039–1.4553	0.8843	8.636	0.88069
ANN	0.0793	0.0358	0.0361	0.0367–0.0588	0.0267	0.261	0.99989
DT	0.1210	0.0742	0.0648	0.0231–0.1066	0.0276	0.468	0.99965
RF	0.1177	0.0699	0.0431	0.0251–0.0589	0.0267	0.311	0.99985
XGBoost	0.1543	0.1142	0.0806	0.0429–0.1066	0.0499	0.582	0.99946

R^2 of 0.88069. R^2 alone provides little discrimination among the nonlinear models because the simulated target covers a wide range and follows smooth deterministic surfaces. RMSE, MAE, and NRMSE provide a clearer comparison.

Figure 3 shows a consistent ranking across the three error metrics. The MLR baseline produced substantially larger errors than all four nonlinear models, with a test RMSE of 1.1957 wax wt.% and NRMSE of 8.636%. In comparison, the nonlinear models achieved NRMSE below 0.6%. This substantial performance gap indicates that the high interpolation accuracy cannot be attributed solely to the smoothness of the simulated response surface and supports the use of nonlinear models for representing the wax-weight response. ANN achieved the lowest test RMSE (0.0361 wax wt.%) and NRMSE (0.261%), followed by RF with an RMSE of 0.0431 wax wt.% and an NRMSE of 0.311%. ANN and RF produced identical MAE values after rounding to four decimal places. Although DT exhibited a higher RMSE, its MAE remained similar, indicating that most predictions were close to the target while a small number of larger errors increased the RMSE. XGBoost produced the largest errors among the evaluated models, although its NRMSE remained below 0.6%. A Friedman test based on curve-level MAE showed no statistically significant difference among ANN, DT, RF, and XGBoost, $\chi^2(3) = 6.60, p = 0.0858$. The planned pairwise Wilcoxon signed-rank tests were therefore not performed.

The cross-validation errors were consistently higher and more variable than the final pressure-held-out errors. Each cross-validation fold contained different complete crude-pressure curves, making some folds inherently more challenging than others. In contrast, the held-out test pressures remained within the pressure range represented during training. The results therefore demonstrate strong interpolation capability across unseen pressure conditions for known crude compositions. The near-unity R^2 values should not be interpreted as evidence of field readiness. Instead, they indicate that the models accurately reproduced a deterministic simulator surface within the represented crude domain. The low normalized errors confirm excellent surrogate accuracy, but they do not address the more demanding question of whether the models generalize to previously unseen crude compositions.

Although ANN achieved the lowest interpolation error, model ranking within the represented crude domain was of secondary importance because the more challenging question was whether the models generalized to previously unseen crude compositions.

4.3 Generalization to Unseen Crude Compositions

LOCO validation changed the interpretation of model quality. Unlike the pressure-held-out evaluation, no model generalized consistently across all five crude compositions. Table 4 summarizes the normalized performance across the held-out crudes. XGBoost achieved the lowest median NRMSE (28.5%), followed by RF (37.0%), ANN (39.1%), and DT (45.0%). However, the mean errors were substantially larger because Tapis produced severe prediction errors for every model. Overall, the

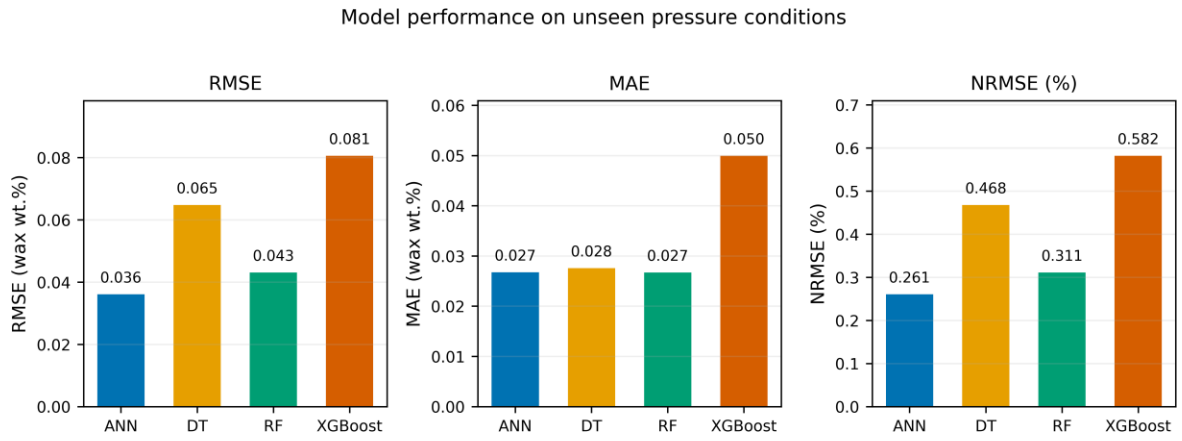


Figure 3 Model Performance on Unseen Pressure Conditions

Table 4 Summary of Leave-One-Crude-Out Generalization Performance

Model	Mean NRMSE (%)	Median NRMSE (%)	NRMSE range (%)	Positive R^2 cases
ANN	57.7	39.1	16.3–112.3	2/5
DT	101.9	45.0	23.7–293.9	2/5
RF	51.6	37.0	20.4–132.3	2/5
XGBoost	49.8	28.5	15.4–133.7	3/5

results indicate that all four models were sensitive to composition domain shift rather than identifying a reliable cross-crude winner.

The marked increase in error relative to the pressure-held-out evaluation demonstrates that composition extrapolation, rather than operating-condition interpolation, is the principal limitation of the developed models.

Figure 4 shows the prediction error for each held-out crude. Tapis was the most difficult case, with NRMSE exceeding 112% for all models and reaching 293.9% for DT. These high normalized errors are partly influenced by the narrow observation wax-weight range of Tapis, since NRMSE was normalized by the range of each held-out crude. A small range therefore amplifies the normalized error relative to its absolute magnitude. Tapis contained no C21+ fraction and lay at the lower boundary of the heavy-end composition range. Crude 3 also produced relatively large errors because it contained the highest C1–C6 fraction (77.065%). Although RF and XGBoost reduced the error for this crude, their NRMSE values remained 37.0% and 46.1%, respectively.

In contrast, crude 4 and crude 5 produced substantially lower errors. Both contained more than 31% C21+ and occupied similar regions of the composition space. When either crude was excluded, the remaining one provided representative training examples. This indicates that cross-crude performance depended primarily on coverage of the composition domain rather than model complexity alone.

The LOCO result is not a conventional overfitting result. The models reproduced the represented simulator surfaces accurately but failed when the input composition moved outside the training domain. This is fundamentally a domain-shift and extrapolation problem rather than a model-capacity problem. Improving generalization therefore requires broader and denser coverage of crude compositions rather than additional model tuning. The LOCO results should therefore be interpreted as cross-crude transfer under fixed model configurations rather than as fully nested estimates of generalization to unseen crude compositions.

4.4 Global and Directional Factor Effects

Temperature was the dominant predictor in the global SHAP ranking. Low temperatures produced positive SHAP values and increased predicted wax weight. High temperatures produced negative values and reduced the prediction. The relation was continuous and monotonic over most of the range. This agrees with the loss of paraffin solubility during cooling.

C1–C6 was the second most important feature by mean absolute SHAP value. The importance bar does not show direction. Figure 5 (c) resolves the interpretation. Low C1–C6 fractions produced positive contributions, while high fractions produced negative contributions. C1–C6 did not promote wax formation. It acted as a composition indicator. Crudes with fewer light ends had higher predicted wax weight.

C21+ showed the expected positive direction at high values. Cases with C21+ above 31% shifted the prediction upward. Cases near 0–3.3% shifted it downward. The relation was not perfectly monotonic because C21+ took only five fixed values and was correlated with the other composition descriptors.

Crude class also received high importance. It should not be interpreted as an independent physical mechanism. The binary class separated the two high-wax crude compositions from the three lower-wax cases. Its contribution reflects grouping information that overlaps with C1-C6 and C21+.

C7–C20 had moderate importance but no clear monotonic direction across the five levels. Pressure had a negligible mean absolute SHAP value. Reporting a separate pressure dependence plot would overemphasise

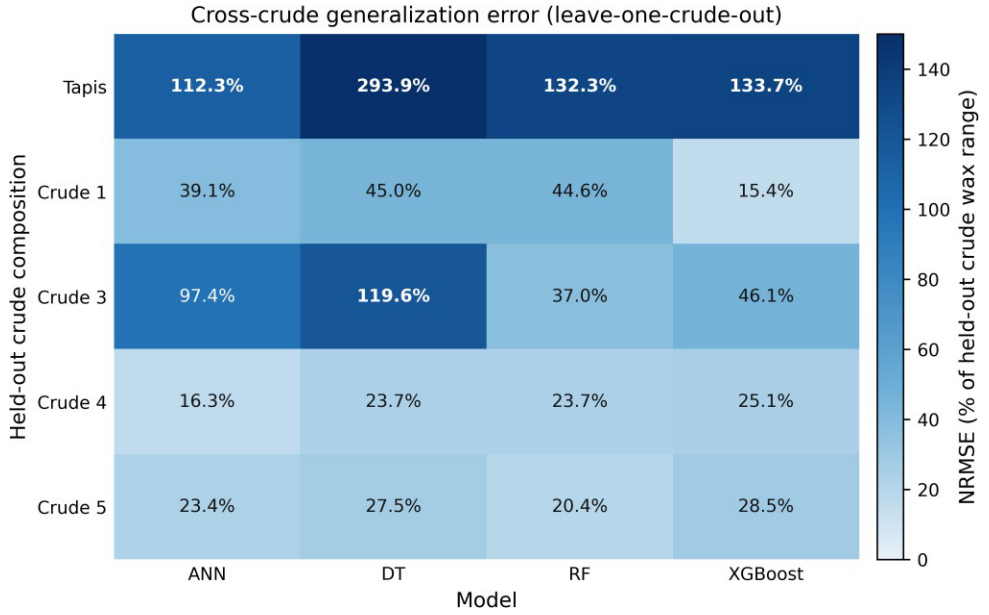


Figure 4 Cross-Crude Generalization Error (Leave-One-Crude-Out)

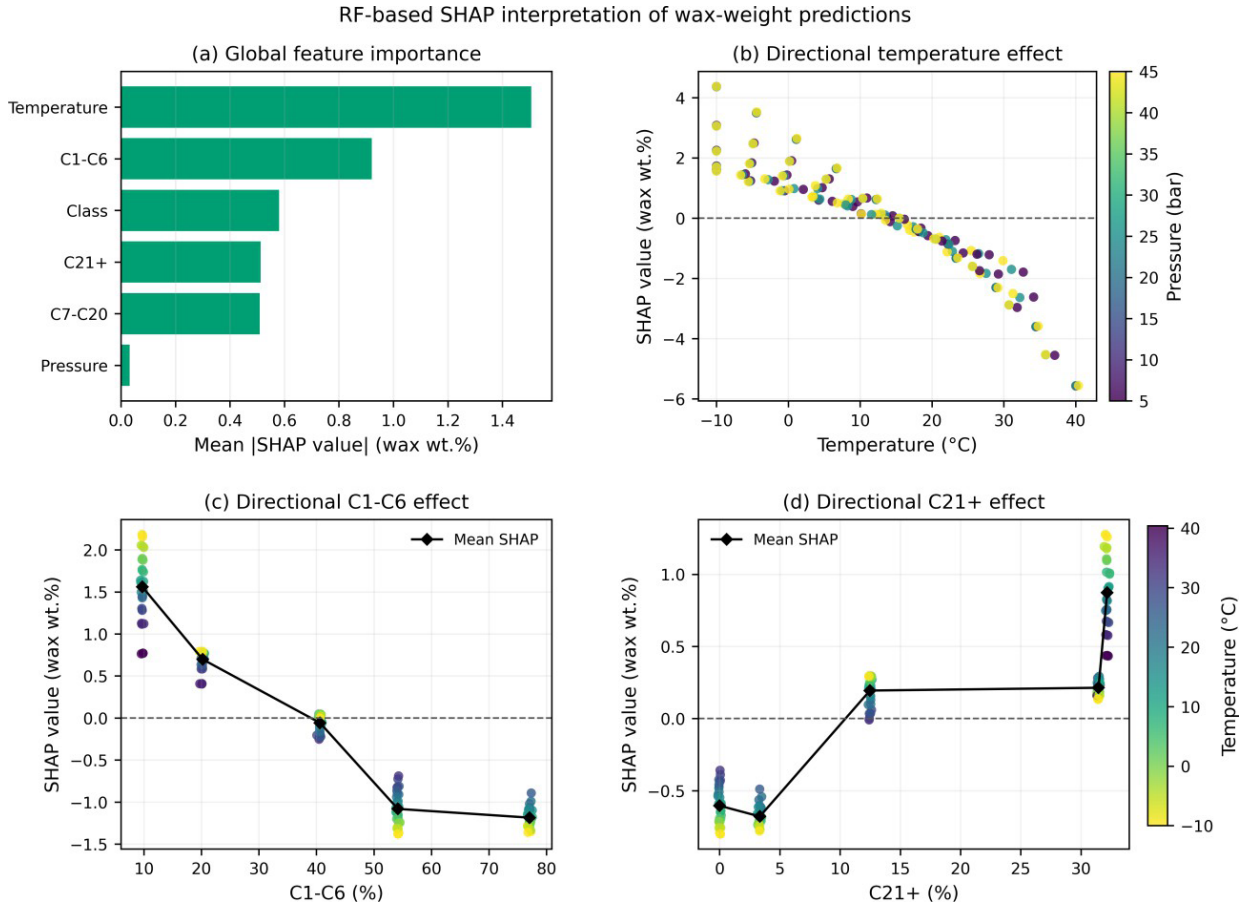


Figure 5 SHAP Analysis of the RF Model

the weak result. The small pressure influence is already visible in Figure 5.

The factor analysis explains the in-domain model but does not remove the cross-crude limitation. The same composition variables that support accurate separation of the five known crudes also make extrapolation difficult when one composition is absent. High importance can therefore indicate both predictive value and dependence on a narrow observation training domain.

5. CONCLUSION

This study developed and compared ANN, DT, RF, and XGBoost models for predicting PVTSim-generated wax precipitation weight using 550 cases from five crude compositions, with MLR included as a baseline. Unlike previous studies, model performance was evaluated using two complementary validation strategies: pressure-held-out testing to assess interpolation within represented crude compositions and LOCO validation to assess generalization to unseen crude compositions.

All four nonlinear models accurately interpolated unseen pressure conditions, with ANN achieving the best performance (RMSE = 0.0361 wax wt.% and NRMSE = 0.261%), demonstrating that machine-learning models can serve as reliable surrogates within the represented PVTSim simulation domain. However, LOCO validation revealed a markedly different outcome. No model generalized consistently across all five crude compositions, indicating that strong in-domain performance does not necessarily translate to reliable prediction for previously unseen crude oils. The results show that composition extrapolation, rather than operating-condition interpolation, is the principal limitation of the developed models.

SHAP analysis further showed that temperature was the dominant predictor, followed by the light- and heavy-end composition fractions. These relationships were physically consistent with established wax precipitation behaviour and explained the successful interpolation within the represented crude domain. At the same time, the analysis demonstrated that the same composition variables responsible for accurate prediction also limited model transferability when new composition patterns were absent from the training data.

Compared with previous studies, this work extends the evaluation of artificial intelligence-based wax prediction by combining multiple model families, structured validation, and model interpretation within a single framework. The results demonstrate that interpretability and predictive generalization should be assessed independently, and that validation strategies should reflect the intended deployment scenario rather than relying solely on random or operating-condition-based data splits.

PVTSim-generated data provide a controlled environment for evaluating operating-condition interpolation and cross-crude generalization, while avoiding the high cost of experimental data generation and limited accessibility of proprietary field data. However, simulator-specific relationships and the absence of experimental noise may contribute to the high in-domain accuracy. The developed models are therefore suitable as fast surrogates for the five represented PVTSim crude compositions, but should not be applied directly to previously unseen crude oils.

Future work should expand the composition space using a larger and more diverse set of crude oils and validate the methodology against experimental wax precipitation measurements. The predicted wax weight could subsequently be integrated with operating and pipeline variables, such as flow rate, pipe geometry, wall temperature, and shear conditions, to develop models for wax deposition or blockage-risk prediction.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Universiti Teknologi PETRONAS for the financial support provided through the Research Fund (YUTP-FRG 015LC0–533).

REFERENCES

Ahmadi, M. A. (2023). Data-driven approaches for predicting wax deposition. *Energy, 265*, 126296. https://doi.org/10.1016/j.energy.2022.126296

Ahmadi, M. (2025). From composition to deposition: A machine learning framework for wax deposition prediction in petroleum fluids using compositional properties. *Physics of Fluids, 37*(7), 072132. https://doi.org/10.1063/5.0276315

Aiyejina, A., Chakrabarti, D. P., Pilgrim, A., & Sastry, M. K. S. (2011). Wax formation in oil pipelines: A critical review. *International Journal of Multiphase Flow, 37*(7), 671–694. https://doi.org/10.1016/j.ijmultiphaseflow.2011.02.007

Alnaimat, F., & Ziauddin, M. (2020). Wax deposition and prediction in petroleum pipelines. *Journal of Petroleum Science and Engineering, 184*, 106385. https://doi.org/10.1016/j.petrol.2019.106385

Amar, N. M., Jahanbani Ghahfarokhi, A., & Ng, C. S. W. (2022). Predicting wax deposition using robust machine learning techniques. *Petroleum, 8*(2), 167–173. https://doi.org/10.1016/j.petlm.2021.07.005

Amiri-Ramsheh, B., Zabihi, R., & Hemmati-Sarapardeh, A. (2023). Modeling wax deposition of crude oils using cascade forward and generalized regression neural networks: Application to crude oil production. *Geoenergy Science and Engineering, 224*, 211613. https://doi.org/10.1016/j.geoen.2023.211613

Behbahani, T. J., Beigi, A. A. M., Taheri, Z., & Ghanbari, B. (2015). Investigation of wax precipitation in crude oil: Experimental and modeling. *Petroleum, 1*(3), 223–230. https://doi.org/10.1016/j.petlm.2015.07.007

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

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

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Huang, Z., Lee, H. S., Senra, M., & Fogler, H. S. (2011). A fundamental model of wax deposition in subsea oil pipelines. *AIChE Journal, 57*(11), 2955–2964. https://doi.org/10.1002/aic.12517

Jalalnezhad, M. J., & Kamali, V. (2016). Development of an intelligent model for wax deposition in oil pipeline. *Journal of Petroleum Exploration and Production Technology, 6*, 129–133. https://doi.org/10.1007/s13202-015-0160-3

Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. *Proceedings of Machine Learning Research, 108*, 2907–2916.

Kamari, A., Khaksar-Manshad, A., Gharagheizi, F., Mohammadi, A. H., & Ashoori, S. (2013). Robust model for the determination of wax deposition in oil systems. *Industrial & Engineering Chemistry Research, 52*(44), 15664–15672. https://doi.org/10.1021/ie402462q

Kim, J., Han, S., Seo, Y., Moon, B., & Lee, Y. (2022). The development of an AI-based model to predict the location and amount of wax in oil pipelines. *Journal of Petroleum Science and Engineering, 209*, 109813. https://doi.org/10.1016/j.petrol.2021.109813

Kiyingi, W., Guo, J. X., Xiong, R. Y., Su, L., Yang, X. H., & Zhang, S. L. (2022). Crude oil wax: A review on formation, experimentation, prediction, and remediation techniques. *Petroleum Science, 19*(5), 2343–2357. https://doi.org/10.1016/j.petsci.2022.08.008

Li, B., Guo, Z., Zheng, L., Shi, E., & Qi, B. (2024). A comprehensive review of wax deposition in crude oil systems: Mechanisms, influencing factors, prediction and inhibition techniques. *Fuel, 357*, 129676. https://doi.org/10.1016/j.fuel.2023.129676

Lira-Galeana, C., Firoozabadi, A., & Prausnitz, J. M. (1996). Thermodynamics of wax precipitation in petroleum mixtures. *AIChE Journal, 42*(1), 239–248. https://doi.org/10.1002/aic.690420120

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Obanijesu, E. O., & Omidiora, E. O. (2008). Artificial neural network's prediction of wax deposition potential of Nigerian crude oil for pipeline safety. *Petroleum Science and Technology, 26*(16), 1977–1991. https://doi.org/10.1080/10916460701399485

Pan, H., Firoozabadi, A., & Fotland, P. (1997). Pressure and composition effect on wax precipitation: Experimental data and model results. *SPE Production & Facilities, 12*(4), 250–258. https://doi.org/10.2118/36740-PA

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Sarkodie, K., Agyepong, J., Agyei, E., Arkoh, G., Arhin, R., Ankomah, F., Adenutsi, C. D., & Erzuah, S. (2026). Machine learning-based surrogate correlation for prediction of wax deposition rate in oil and gas pipelines. *Petroleum, 12*(2), 350–365. https://doi.org/10.1016/j.petlm.2026.03.002

Singh, P., Venkatesan, R., Fogler, H. S., & Nagarajan, N. (2000). Formation and aging of incipient thin film wax-oil gels. *AIChE Journal, 46*(5), 1059–1074. https://doi.org/10.1002/aic.690460517

Xie, Y., & Xing, Y. (2017). A prediction method for the wax deposition rate based on a radial basis function neural network. *Petroleum, 3*(2), 237–241. https://doi.org/10.1016/j.petlm.2016.08.003

Yao, B., Chen, J., Li, C., Yang, F., Sun, G., & Lu, Y. (2021). Prediction of wax deposits for crude pipelines using time-dependent data mining. *SPE Journal*, *26*(4). https://doi.org/10.2118/205374-PA