Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i2.455

Research Article

Comparative Evaluation of Machine Learning Models for Heavy Crude Oil Viscosity Prediction Using Repeated Nested Cross-Validation and Independent Holdout Testing

Enggie Hendrawan Saputra[1]; Ilham Ari Elbaith Zaeni [2]; Didik Dwi Prasetya [3]; Azlan Mohd Zain[4]; Welly Antonius[5]; I Made Wirawan[6]

[1] Universitas Negeri Malang, Kota Malang, 65115, Indonesia, enggie.hendrawan.2405348@students.um.ac.id
[2] Universitas Negeri Malang, Kota Malang, 65115, Indonesia, ilham.ari.ft@um.ac.id
[3] Universitas Negeri Malang, Kota Malang, 65115, Indonesia, didikdwi@um.ac.id
[4] Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia, azlanmz@utm.my
[5] PT Kilang Pertamina Internasional, Daerah Khusus Ibukota Jakarta 10110, Indonesia, wellysiku5@gmail.com
[6] Universitas Negeri Malang, Kota Malang, 65115, Indonesia, made.wirawan.ft@um.ac.id

Correspondence author

Received 29 March 2026 Accepted 19 May 2026; Published 31 July 2026

Abstract:
Introduction: Accurate prediction of heavy crude oil viscosity is important for reservoir engineering, production planning, and flow assurance because viscosity strongly affects fluid mobility and transport behavior. This study comparatively evaluates established machine learning models under a rigorous validation protocol rather than proposing a new predictive framework. **Method**: A published Middle Eastern heavy crude-oil dataset containing 196 development measurements and 47 independent holdout measurements was used. Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, and the Beggs–Robinson correlation were evaluated using repeated nested cross-validation with five outer folds repeated twice and five inner folds. Preprocessing and hyperparameter selection were embedded within the validation pipeline, while the untouched holdout set was used only for final evaluation. **Results and Discussion**: Gradient Boosting achieved the best internal performance with R^2 = 0.99313 and RMSE = 11.41 cP. On the independent holdout set, it achieved R^2 = 0.99308, RMSE = 8.43 cP, MAE = 6.64 cP, and MAPE = 0.78%, outperforming Random Forest and Support Vector Regression. Residual diagnostics showed no detectable heteroscedasticity, while permutation importance identified temperature and C7+ as the dominant predictors. **Conclusion**: Gradient Boosting provides highly accurate viscosity predictions within the sampled domain; however, the absence of row-level oil identifiers and external reservoir data limits conclusions regarding oil-disjoint and field-level generalization.

Keywords: Heavy Crude Oil Viscosity; Machine Learning; Gradient Boosting; Repeated Nested Cross-Validation; Independent Holdout Testing; Comparative Evaluation.

1. Introduction

Accurate estimation of crude oil viscosity is central to reservoir simulation, production planning, enhanced oil recovery, and pipeline design because viscosity governs fluid mobility, pressure loss, and transport behavior [1], [2], [3], [4], [5]. Heavy-oil viscosity is strongly nonlinear with respect to temperature and composition, particularly the

heavy C7+ fraction, which limits the transferability of simple correlations across different crude-oil systems [5], [6], [7].

Machine-learning models have therefore been investigated as nonlinear alternatives to laboratory-intensive measurements and empirical correlations. Prior studies have reported strong results using support-vector models, neural networks, decision trees, and boosting methods [1], [3], [4], [9], [12]. However, reported performance is difficult to compare because dataset sizes, measurement dependence, validation strategies, baselines, and test domains differ substantially. Random record-level splits may also overstate generalization when multiple measurements originate from the same crude oil.

The contribution of this study is a rigorous comparative evaluation rather than a new machine-learning framework. Established tools—nested cross-validation, corrected statistical testing, residual diagnostics, and permutation importance—are combined within one reproducible protocol to compare linear, kernel-based, bagging, and boosting regressors on a limited heavy-crude dataset. The revision also restores the independent 47-record validation sheet that was not used in the original analysis code.

Table 1 positions the present study against representative viscosity-prediction studies. The comparison distinguishes the number of measurements from the number of independent crude-oil samples and highlights whether model assessment used a separate validation set or only internal resampling.

Table 1. Comparison of Representative Heavy Crude Oil Viscosity Prediction Studies

Study	Data structure	Validation	Models / baseline	Reported result	Main limitation for comparison
Oloso et al. [1]	About 250 dead/saturated-oil records and 910 undersaturated-oil records	Single random 70/30 split	Ensemble SVR and empirical correlations	Ensemble SVR improved the selected error criteria	No nested or group-aware validation reported
Li et al. [9]	28 Middle Eastern heavy oils; 196 development measurements at 20–80 °C; 47 additional measurements	Development set plus additional validation measurements	DT, MLP, and GRNN	Best model: MLP, $R^2 = 0.978$	Oil-level separation in model development should be stated explicitly
Langeroudy et al. [12]	1,368 Iranian reservoir-condition records	Random 80/20 split and 10-fold CV on training data	XGBoost, CatBoost, and Gradient Boosting; comparisons with established approaches	Best model: $R^2 = 0.9976$; AARD = 1.968%	Random record-level splitting may not represent reservoir-level transfer
Present study	28 reported crude-oil samples; 196 development measurements; 47 independent validation measurements	Repeated 5-fold nested CV (2 repeats) plus untouched 47-record holdout	Beggs–Robinson, Linear Regression, SVR, RF, and GB	GB holdout: $R^2 = 0.9931$; RMSE = 8.43 cP	Row-level oil/reservoir IDs unavailable; no external reservoir test

Accordingly, this study aims to: (1) compare Linear Regression, Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting (GB) under the same leakage-controlled protocol; (2) benchmark the models against the Beggs–Robinson empirical correlation; (3) estimate internal performance using repeated nested cross-validation and test the final models once on an untouched 47-record holdout set; (4) quantify uncertainty using confidence intervals and dependence-corrected comparisons; and (5) interpret residual behavior and held-out permutation importance without claiming reservoir-level generalization.

2. Method

This study uses an experimental approach with systematic stages to develop and test a website-based boycott product detection system using the Convolutional Neural Network (CNN) method. The initial steps taken are data collection, dataset, data split, data preprocessing, algorithm application, evaluation, and application testing. Details of the method stages can be seen in **Figure 1**.

Research Design

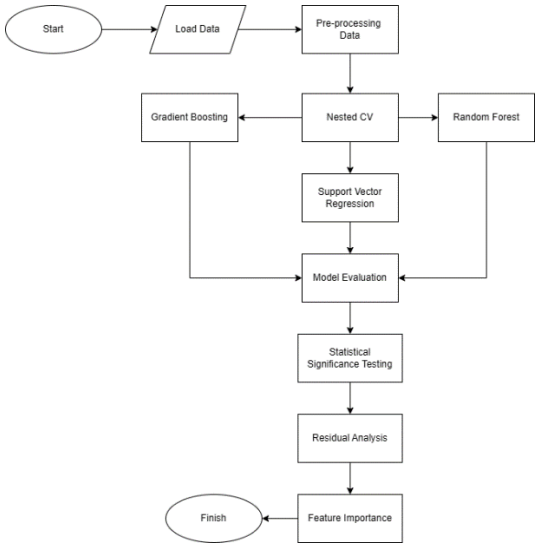

Figure 1. Research Design

The revised workflow consists of data provenance verification, data-integrity auditing, repeated nested cross-validation on the 196-record development sheet, final hyperparameter selection, one-time evaluation on the independent 47-record holdout sheet, corrected statistical comparisons, residual diagnostics, extreme-value sensitivity analysis, and held-out permutation importance. The holdout sheet is excluded from preprocessing selection, hyperparameter tuning, and model ranking during internal validation.

All data-dependent preprocessing is embedded in scikit-learn pipelines. Median imputation is fitted within each training fold for every model, while Min–Max scaling is additionally fitted within each training fold for SVR. Thus, no imputation statistic, scaling parameter, or hyperparameter is learned from an outer-test fold or from the independent holdout data.

The internal assessment uses five outer folds repeated twice, producing ten paired outer-split scores, with five-fold inner cross-validation for hyperparameter selection. Because these scores remain statistically dependent, pairwise RMSE differences are evaluated using the Nadeau–Bengio corrected test, 95% confidence intervals, standardized paired effect sizes, and Holm adjustment for multiple comparisons [14], [15], [20].

Data Description

The workbook reproduces the published dataset described by Kamel et al. [8] and Li et al. [9]. The source study reports 28 heavy crude-oil samples collected from Middle Eastern oil fields, with 196 measurements used for model development and 47 additional measurements used for validation. The predictors are API gravity, temperature (°C), and the C1, C2, C3, C4–C6, and C7+ compositional fractions; the target is viscosity in centipoise (cP). In the development sheet, temperature ranges from 20 to 80 °C and viscosity ranges from 632.88 to 1267.65 cP (median 850.255 cP). The holdout sheet covers 20–80 °C and 660.32–1080.59 cP.

Let the dataset be defined as:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

Where x_i is the seven-variable predictor vector for measurement i and y_i is its measured viscosity. Although the source publication reports 28 crude-oil samples, the supplied workbook does not contain a row-level oil, field, or reservoir identifier. Therefore, the dependence structure among repeated measurements cannot be reconstructed, and group-disjoint splitting is not claimed in this study.

The audit found no missing cells, no exact duplicate rows, no duplicate predictor vectors within either sheet, and no exact predictor-vector overlap between the development and holdout sheets. Hydrocarbon fractions were retained as reported; their partial sums are below 100% because the available columns do not represent a complete compositional assay. Median imputation remains inside the pipelines for reproducibility even though no missing values were observed.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2)$$

No record was removed solely because of an extreme viscosity value. Sensitivity was evaluated by repeating holdout performance calculations after excluding observations outside the 5th–95th percentile range of the development target. This analysis tests whether the main conclusion is disproportionately driven by the most extreme viscosity observations.

Data Integrity and Measurement Dependence

Table 2 summarizes the dataset structure, measurement domain, and leakage audit. The missing row-to-oil identifier is treated as a substantive limitation: the 47-record sheet provides an independent published holdout at the record level, but neither oil-level nor reservoir-level independence can be verified from the supplied file.

Table 2. Dataset Structure, Measurement Domain, and Data Leakage Audit

Item	Required report	Value
Dataset provenance	Owner/source, access status, and citation	Published Middle Eastern heavy-crude dataset from Kamel et al. [8], reproduced by Li et al. [9]
Independent units	Number of distinct oils and reservoirs/fields	28 crude-oil samples reported by source; fields/reservoir count not reported in workbook
Repeated measurements	Records per oil and conditions varied	196 development and 47 validation measurements; row-to-oil mapping unavailable
Viscosity domain	Minimum, median, maximum, and units	Development: 632.88–1267.65 cP (median 850.255); holdout: 660.32–1080.59 cP
Temperature domain	Minimum, median, maximum, and units	20–80 °C in both sheets; development median 50 °C
Missing data	Count and percentage by feature	0 missing cells in both sheets
Duplicates	Exact duplicates and repeated predictor rows	0 exact duplicate rows; 0 duplicate predictor vectors
Compositional checks	Range and sum of hydrocarbon fractions	Reported partial fractions retained; sums are below 100% because only selected fractions are available

Item	Required report	Value
Group variable	Oil/reservoir identifier used for splitting	Not present; oil-/reservoir-disjoint splitting cannot be reconstructed

Nested Cross-Validation Framework

Model assessment uses repeated record-level nested cross-validation on the 196-record development sheet: five outer folds repeated twice and five inner folds. The outer splits are generated with shuffling and random seed 42. For each outer split, the corresponding training records are used for all preprocessing and hyperparameter selection, while the outer-test records remain untouched until final scoring. The 47-record holdout sheet is not used in this process.

For each repetition r and outer fold k, the record sets satisfy, let D be divided into K outer folds:

$$D = \bigcup_{k=1}^{K} D_k \tag{9}$$

For each outer fold:

1. One record-level fold is held out as the outer-test partition.
2. The remaining records enter five-fold inner cross-validation, where the complete preprocessing pipeline and model hyperparameters are selected by minimizing RMSE.
3. The selected pipeline is refitted on the complete outer-training partition and evaluated once on the untouched outer-test fold.

Preprocessing Pipelines, Hyperparameter Search, and Reproducibility

The exact search spaces are reported in Table 3. GridSearchCV minimizes inner-fold RMSE. Linear Regression has no tuned hyperparameters. SVR uses median imputation, Min–Max scaling, and an RBF kernel. RF and GB use median imputation without scaling. The empirical Beggs–Robinson correlation is evaluated without learned preprocessing and is interpreted only as a historical baseline because its calibration domain differs from the present heavy-oil measurements [28].

Table 3. Preprocessing Pipelines and Hyperparameter Search Spaces

Model	Pipeline	Candidate hyperparameters for the rerun
Linear Regression	Median imputer → Linear Regression	No tuned hyperparameters
SVR (RBF)	Median imputer → Min–Max scaler → SVR	C: {10, 100}; gamma: {0.01, 0.1}; epsilon: {0.1, 10}
Random Forest	Median imputer → RF	n_estimators: {200}; max_depth: {None, 8}; min_samples_leaf: {1, 2}; max_features: {0.8, 1.0}; bootstrap: {True}
Gradient Boosting	Median imputer → GB	n_estimators: {200, 400}; learning_rate: {0.03, 0.05}; max_depth: {2, 3}; min_samples_leaf: {1}; subsample: {0.8}
Beggs–Robinson	No learned preprocessing	Original dead-oil correlation using API gravity and temperature [28]

The analysis uses Python 3.12.13, scikit-learn 1.6.1, NumPy 2.0.2, SciPy 1.16.3, pandas 2.2.2, and statsmodels 0.14.6 on Linux. Random seed 42 is used for fold generation, model stochasticity, bootstrap resampling, and permutation importance. The complete executable script, selected parameters, split-level metrics, predictions, and software environment are provided as supplementary reproducibility outputs.

Regression Models

Three regression paradigms were evaluated to capture nonlinear relationships between compositional features and viscosity [21], [22], [23].

1. Support Vector Regression (SVR)

SVR aims to find a function $f(x)$ that deviates from the observed targets by no more than ϵ while maintaining maximum flatness [24], [25]. The optimization problem is formulated as:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{3}$$

Subject to:

$$\begin{aligned} y_i - (\omega \cdot \phi(x_i) + b) &\leq \in + \xi_i \\ (\omega \cdot \phi(x_i) + b) - y_i &\leq \in + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \tag{4}$$

Where $\phi(x)$ maps the input space to a high-dimensional feature space and C controls the trade-off between model complexity and error tolerance. The radial basis function (RBF) kernel was used:

$$K(x_i, x_j) = \exp(-\gamma \left|\left|x_i - x_j\right|\right|^2) \tag{5}$$

2. Random Forest (RF)

Random Forest is an ensemble of decision trees trained using bootstrap aggregation (bagging) [26], [27]. The prediction is obtained by averaging individual tree predictions:

$$\hat{y} = \frac{1}{T}\sum_{i=1}^{T} f_t(x) \tag{6}$$

Where T is the number of trees and $f_t(x)$ is the prediction from tree t. Random feature selection at each split introduces decorrelation between trees, reducing variance.

3. Gradient Boosting (GB)

Gradient Boosting builds an additive model sequentially by minimizing a differentiable loss function [28], [29]. At iteration m, the model is updated as:

$$F_m(x) = F_{m-1}(x) + vh_m(x) \tag{7}$$

Where $h_m(x)$ is the weak learner trained on the negative gradient of the loss function, and v is the learning rate. For regression with squared error loss:

$$L = \sum_{i=1}^{N}\left(y_i - F(x_i)\right)^2 \tag{8}$$

The negative gradient corresponds to residual errors, allowing the model to iteratively reduce bias.

Performance Metrics

Performance is summarized using R^2, RMSE, MAE, and MAPE. The minimum viscosity is 632.88 cP in the development set and 660.32 cP in the holdout set, so MAPE is numerically stable in this application. RMSE in cP is used as the primary model-selection and pairwise-comparison metric because it preserves the engineering unit and penalizes large prediction errors.

- Coefficient of Determination (R^2).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{10}$$

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{11}$$

- Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{12}$$

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{13}$$

Statistical Significance Testing

To determine whether performance differences between models were statistically significant, paired t-tests were conducted across outer cross-validation folds. For two models A and B with fold-wise scores s_A^k and s_B^k:

$$t = \frac{\bar{d}}{s_d/\sqrt{K}} \tag{14}$$

Where $d_k = s_A^k - s_B^k, \bar{d}$ is the mean difference, and s_d is the standard deviation of differences.

The null hypothesis assumes no significant difference between models. A significance level of $\alpha = 0.05$ was used.

Residual Diagnostics

Residuals were computed as:

$$e_i = y_i - \hat{y}_i \tag{15}$$

Residual diagnostics are calculated exclusively from the independent holdout predictions. For each model, residual mean and standard deviation are reported together with the Breusch–Pagan test and the Spearman association between absolute residual magnitude and fitted viscosity. Because the holdout contains only 47 records, a non-significant diagnostic is interpreted as no detectable heteroscedasticity rather than proof of constant variance.

Permutation Feature Importance

Permutation importance is computed only after model selection, using the untouched 47-record holdout set. Each feature is permuted 50 times and importance is expressed as the increase in holdout RMSE relative to the unpermuted prediction. The mean, standard deviation, and empirical 95% interval across permutations are reported [16], [26], [27]:

$$FI_j = \mathbb{E}[L(f(X), y)] - \mathbb{E}\left[L(f(X_{perm(j)}), y)\right] \quad (15)$$

Where $X_{perm(j)}$ denotes the dataset with feature j randomly permuted.

This method provides a model-agnostic estimate of feature contribution

3. Result and Discussion

Result

Across ten repeated outer splits, GB achieved the strongest internal performance, with R^2 = 0.99313 (95% CI: 0.99077–0.99549), RMSE = 11.41 cP (10.03–12.80), MAE = 8.94 cP, and MAPE = 1.03%. RF ranked second (R^2 = 0.97570; RMSE = 21.47 cP), followed by Linear Regression and SVR. The Beggs–Robinson correlation produced very large errors, indicating that it is not transferable to this dataset without recalibration.

The data audit identified 196 development records and 47 holdout records, seven predictors, zero missing cells, zero exact duplicate rows, zero duplicate predictor vectors, and zero exact predictor overlap across sheets. The source reports 28 crude-oil samples, but the workbook does not map records to those oils. Consequently, the effective number of independent units and the possibility of repeated-oil dependence remain unresolved:

Table 4. Repeated Record-Level Nested Cross-Validation Performance (10 Outer Splits)

Model	R^2 (mean; 95% CI)	RMSE (mean; 95% CI)	MAE (mean; 95% CI)	MAPE (mean; 95% CI)
Linear Regression	0.88540 (0.87573–0.89507)	47.99 (45.49–50.49)	41.61 (39.34–43.89)	4.63% (4.42–4.83)
Beggs–Robinson*	−215.16085 (−341.08708–−89.23463)	1914.56 (1250.36–2578.76)	1080.46 (837.31–1323.61)	116.37% (94.96–137.78)
SVR (RBF)	0.81730 (0.78986–0.84474)	60.76 (53.70–67.82)	39.20 (33.72–44.68)	4.02% (3.54–4.49)
Random Forest	0.97570 (0.96822–0.98319)	21.47 (18.53–24.41)	16.96 (14.73–19.18)	1.96% (1.71–2.21)
Gradient Boosting	0.99313 (0.99077–0.99549)	11.41 (10.03–12.80)	8.94 (7.98–9.91)	1.03% (0.91–1.14)

Corrected pairwise comparisons are shown in **Table 5**. GB reduced mean outer-split RMSE by 10.05 cP relative to RF (95% corrected CI: 5.55–14.56; Holm-adjusted p = 0.0042; paired effect size d_z = 2.99) and by 49.35 cP relative to SVR (34.89–63.80; adjusted p = 0.00023; d_z = 4.57). RF also outperformed SVR by 39.29 cP (24.12–54.47; adjusted p = 0.0017; d_z = 3.47).

Table 5. Independent Holdout Performance with Bootstrap 95% Confidence Intervals

Model Comparison	Paired t-test p-value	Wilcoxon p-value
SVR vs RF	39.29 cP (24.12–54.47)	0.0017
SVR vs GB	49.35 cP (34.89–63.80)	0.00023
RF vs GB	10.05 cP (5.55–14.56)	0.0042

Table 6. Corrected Pairwise Comparisons of Outer-Split Performance

Model	R^2 (95% CI)	RMSE, cP (95% CI)	MAE, cP (95% CI)	MAPE (95% CI)
Linear Regression	0.82590 (0.74732–0.86710)	42.28 (36.38–47.79)	37.40 (31.72–43.24)	4.38% (3.70–5.08)
Beggs–Robinson*	−55.20804 (−81.88909–−41.32952)	759.69 (722.56–792.92)	743.31 (693.38–784.89)	87.54% (81.21–92.37)
SVR (RBF)	0.91748 (0.84644–0.96504)	29.11 (18.77–41.34)	20.57 (15.77–27.18)	2.39% (1.87–3.05)
Random Forest	0.96091 (0.93241–0.97843)	20.04 (14.97–25.05)	15.40 (11.89–19.39)	1.79% (1.39–2.24)
Gradient Boosting	0.99308 (0.98772–0.99637)	8.43 (6.43–10.20)	6.64 (5.22–8.14)	0.78% (0.61–0.97)

The identical Wilcoxon p-value of 0.0625 from the original five-fold analysis was a consequence of its extremely low discrete resolution and is no longer used. On the untouched 47-record holdout set (Table 6), GB achieved R^2 = 0.99308 (bootstrap 95% CI: 0.98772–0.99637), RMSE = 8.43 cP (6.43–10.20), MAE = 6.64 cP (5.22–8.14), and MAPE = 0.78% (0.61–0.97%). RF achieved R^2 = 0.96091 and RMSE = 20.04 cP, while SVR achieved R^2 = 0.91748 and RMSE = 29.11 cP.

Figures 2–4 present actual-versus-predicted values on the independent holdout set, and Figures 5–7 present the corresponding residuals. All labels were regenerated at publication resolution. Figure 8 reports permutation importance calculated exclusively on the independent holdout set.

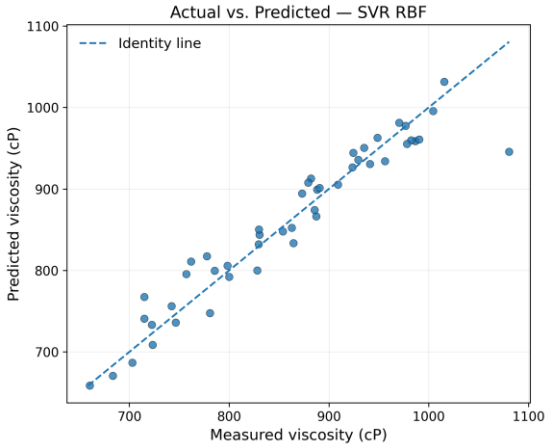

Figure 2. Actual vs Predicted – Support Vector Regression

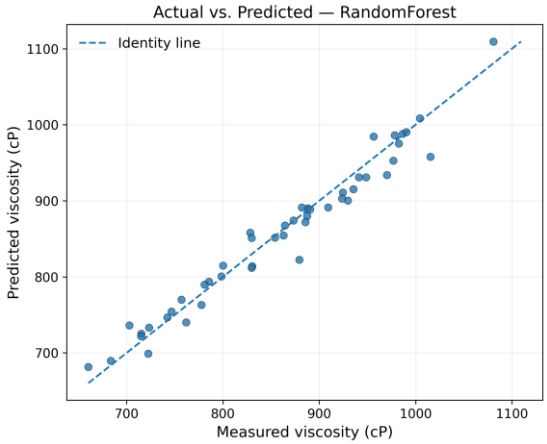

Figure 3. Actual vs Predicted – Random Forest

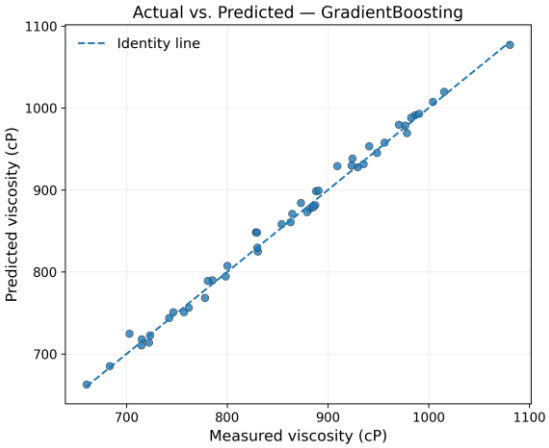

Figure 4. Actual vs Predicted – Gradient Boosting

The holdout plots show progressively tighter agreement from SVR to RF and GB. GB predictions lie closest to the identity line across the observed range, with no single isolated viscosity extreme accounting for the overall fit. Excluding the two holdout observations outside the development-set 5th–95th target percentiles changed GB only slightly (RMSE 8.43 to 8.60 cP; R^2 0.99308 to 0.99185).

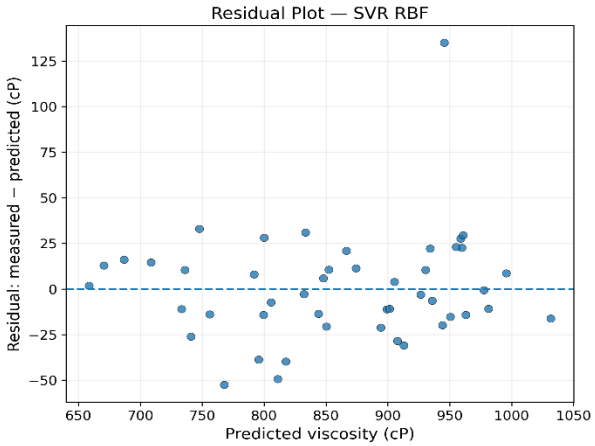

Figure 5. Residual Plot – Support Vector Regression

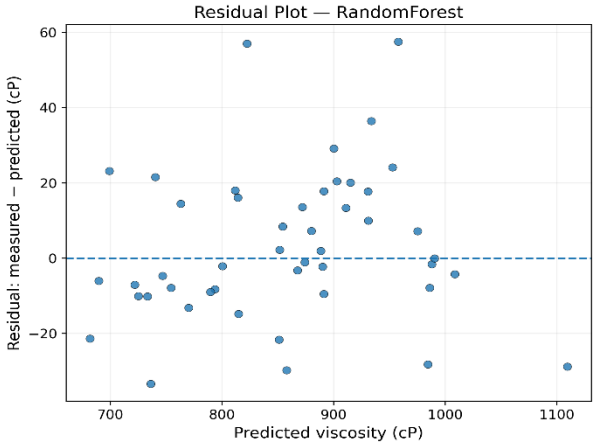

Figure 6. Residual Plot – Random Forest

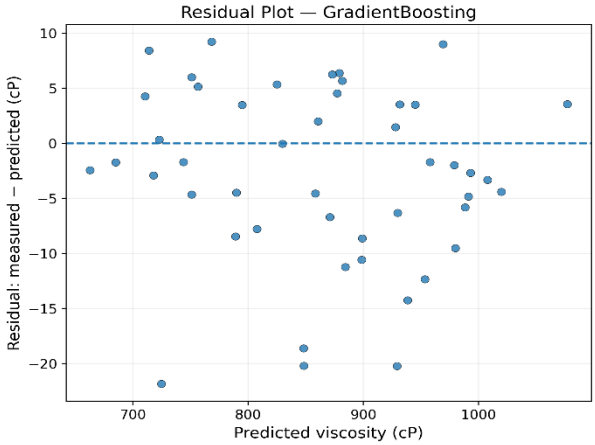

Figure 7. Residual Plot – Gradient Boosting

On the holdout set, GB residuals had a mean of −2.90 cP and a standard deviation of 8.00 cP. The Breusch–Pagan test did not detect heteroscedasticity for GB ($p = 0.860$), and absolute residual magnitude was not associated with fitted viscosity (Spearman $\rho = 0.075$, $p = 0.617$). RF and SVR also had non-significant Breusch–Pagan results ($p = 0.415$ and 0.495). Given $n = 47$, these findings indicate no detectable variance pattern rather than definitive homoscedasticity.

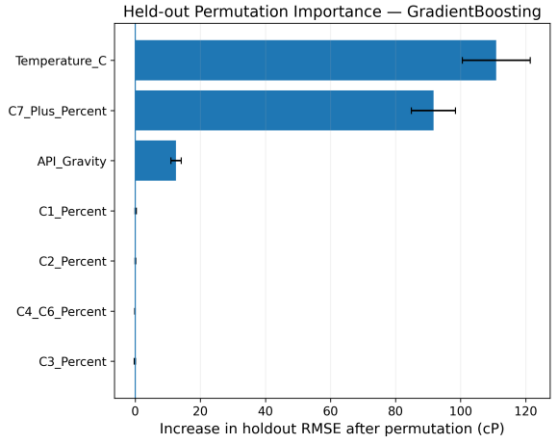

Figure 8. Held-out Permutation Importance – Gradient Boosting

Held-out permutation importance identified temperature as the strongest predictor, with a mean RMSE increase of 110.97 cP after permutation (95% permutation interval: 92.78–127.28), followed by C7+ at 91.65 cP (81.16–102.81) and API gravity at 12.54 cP (9.15–14.74). The intervals for C1, C2, C3, and C4–C6 included zero. Because these compositional variables are correlated, the near-zero marginal values should not be interpreted as proof that the corresponding fractions are physically irrelevant.

Discussion

The revised analysis supports GB as the best-performing model within the sampled measurement domain, but it does not establish a universally robust framework. Its advantage is observed consistently under repeated record-level nested CV and on the separate published holdout sheet. The linear baseline performs substantially worse than GB, confirming that nonlinear interactions among temperature, API gravity, and composition are practically important in this dataset.

The holdout GB RMSE of 8.43 cP corresponds to less than 1% of the holdout mean viscosity, and the MAE is 6.64 cP. These errors are encouraging for interpolation within the represented temperature and compositional ranges. Nevertheless, practical superiority for reservoir engineering requires comparison with laboratory repeatability, deployment tolerances, and genuinely external oil fields or reservoirs. The failure of the unmodified Beggs–Robinson correlation further emphasizes that empirical and machine-learning models must be evaluated within their calibration domains.

4. Conclusion

This study presents a rigorous comparative evaluation of empirical, linear, kernel-based, bagging, and boosting approaches for heavy crude-oil viscosity prediction. Using leakage-controlled repeated nested cross-validation on 196 development measurements, GB achieved $R^2 = 0.99313$ and RMSE = 11.41 cP. On the untouched 47-record holdout set, it achieved $R^2 = 0.99308$, RMSE = 8.43 cP, MAE = 6.64 cP, and MAPE = 0.78%. Corrected pairwise tests with Holm adjustment confirmed lower internal RMSE than RF and SVR. Held-out residual diagnostics and extreme-value sensitivity analyses did not reveal an obvious instability, while temperature and C7+ were the dominant predictive variables.

The results should be regarded as encouraging rather than definitive. The source reports 28 crude-oil samples, but the supplied workbook lacks row-level oil and reservoir identifiers; dependence among repeated measurements and oil-disjoint generalization therefore cannot be evaluated. The dataset is small, the 47-record holdout is drawn from the same published data source, no independent reservoir or field dataset is available, and statistical inference remains based on fold-level resampling. Future work should release sample identifiers, apply group-disjoint validation, evaluate genuinely external reservoirs, quantify measurement uncertainty, and examine conditional or grouped importance for correlated compositional predictors.

References:

[1] M. A. Oloso, M. G. Hassan, M. B. Bader-El-Den, and J. M. Buick, “Ensemble SVM for characterisation of crude oil viscosity,” *J. Pet. Explor. Prod. Technol.*, vol. 8, no. 2, pp. 531–546, Jun. 2018, doi: https://doi.org/10.1007/s13202-017-0355-x.

[2] F. Souas, A. Safri, and A. Benmounah, “A review on the rheology of heavy crude oil for pipeline transportation,” *Pet. Res.*, vol. 6, no. 2, pp. 116–136, Jun. 2021, doi: https://doi.org/10.1016/j.ptlrs.2020.11.001.

[3] E. Bahonar, M. Chahardowli, Y. Ghalenoei, and M. Simjoo, “New correlations to predict oil viscosity using data mining techniques,” *J. Pet. Sci. Eng.*, vol. 208, p. 109736, Jan. 2022, doi: https://doi.org/10.1016/j.petrol.2021.109736.

[4] U. Sinha, B. Dindoruk, and M. Soliman, “Machine learning augmented dead oil viscosity model for all oil types,” *J. Pet. Sci. Eng.*, vol. 195, p. 107603, Dec. 2020, doi: https://doi.org/10.1016/j.petrol.2020.107603.

[5] M. Ghanavati, M.-J. Shojaei, and A. R. S. A., "Effects of Asphaltene Content and Temperature on Viscosity of Iranian Heavy Crude Oil: Experimental and Modeling Study," *Energy \& Fuels*, vol. 27, no. 12, pp. 7217–7232, Dec. 2013, doi: https://doi.org/10.1021/ef400776h.

[6] H. S. Naji, "Comparative Study of the C7+ Characterization Methods: An Object-Oriented Approach," *Arab. J. Sci. Eng.*, vol. 36, no. 7, pp. 1423–1446, Nov. 2011, doi: https://doi.org/10.1007/s13369-011-0118-9.

[7] C. H. Whitson, "Effect of C7+ Properties on Equation-of-State Predictions," *Soc. Pet. Eng. J.*, vol. 24, no. 06, pp. 685–696, Dec. 1984, doi: https://doi.org/10.2118/11200-PA.

[8] M. Baghani, "Energy optimization in electrochemical oxidation for wastewater treatment using interpretable machine learning," *Chem. Eng. J. Adv.*, vol. 25, 2026, doi: https://doi.org/10.1016/j.ceja.2026.101044.

[9] Y. Bouslihim, "Predicting soil organic matter from color indices: economic and technical feasibility in semi-arid agricultural soils," *Carbon Res.*, vol. 5, no. 1, 2026, doi: https://doi.org/10.1007/s44246-025-00240-6.

[10] O. Alomair, E. Folad, and A. Elsharkawy, "Effects of asphaltene and resin contents on crude oil viscosity: Experimental and modeling studies," *Fuel*, vol. 405, p. 136575, 2026, doi: https://doi.org/10.1016/j.fuel.2025.136575.

[11] Y. He *et al.*, "rediction of High-Pressure Physical Properties of crude oil Using Explainable Machine Learning Models," *IEEE Access*, vol. 13, pp. 4912–4931, 2025, doi: https://doi.org/10.1109/ACCESS.2024.3522595.

[12] K. Peiro Ahmady Langeroudy, P. Kharazi Esfahani, and M. R. Khorsand Movaghar, "Enhanced intelligent approach for determination of crude oil viscosity at reservoir conditions," *Sci. Rep.*, vol. 13, no. 1, p. 1666, Jan. 2023, doi: https://doi.org/10.1038/s41598-023-28770-2.

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, Oct. 2001, doi: https://doi.org/10.1214/aos/1013203451.

[14] C. Nadeau and Y. Bengio, "Inference for the {Generalization} {Error}," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003, doi: https://doi.org/10.1023/A:1024068626366.

[15] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: https://doi.org/10.1162/089976698300017197.

[16] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, Dec. 2008, doi: https://doi.org/10.1186/1471-2105-9-307.

[17] B. Dou, "Machine Learning Methods for Small Data Challenges in Molecular Science," *Chemical Reviews*, vol. 123, no. 13. pp. 8736–8780, 2023, doi: https://doi.org/10.1021/acs.chemrev.3c00189.

[18] P. T. Zakowicz, "Machine learning helps predict early onset psychosis with serum protein biomarkers, neuropsychometry, and clinicodemographic data," *Sci. Rep.*, vol. 16, no. 1, 2026, doi: https://doi.org/10.1038/s41598-025-33765-2.

[19] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: https://doi.org/10.1109/ASYU50717.2020.9259880.

[20] Y. Nie, "Deep Melanoma classification with K-Fold Cross-Validation for Process optimization," *IEEE Med. Meas. Appl. MeMeA 2020 - Conf. Proc.*, 2020, doi: https://doi.org/10.1109/MeMeA49120.2020.9137222.

[21] J. M. Dahr, "A Hard Voting Ensemble Model of the Logistic Regression, Support Vector Machine and Random Forest for Network Intrusion Detection," *Inform. Slov.*, vol. 49, no. 27, pp. 43–56, 2025, doi: https://doi.org/10.31449/inf.v49i27.8573.

[22] Z. Liu, "Prediction of landfill gases concentration based on Grey Wolf Optimization – Support Vector Regression during landfill excavation process," *Waste Manag.*, vol. 198, pp. 128–136, 2025, doi: https://doi.org/10.1016/j.wasman.2025.02.040.

[23] J. C. M. Sánchez, "Improving wheat yield prediction through variable selection using Support Vector Regression, Random Forest, and Extreme Gradient Boosting," *Smart Agricultural Technology*, vol. 10.. 2025, doi: https://doi.org/10.1016/j.atech.2025.100791.

[24] C. C. Onyekwena, "Support vector machine regression to predict gas diffusion coefficient of biochar-amended soil," *Appl. Soft Comput.*, vol. 127, 2022, doi: https://doi.org/10.1016/j.asoc.2022.109345.

[25] H. Su, "Support Vector Regression-Based Reduced- Reference Perceptual Quality Model for Compressed Point Clouds," *IEEE Trans. Multimed.*, vol. 26, pp. 6238–6249, 2024, doi: https://doi.org/10.1109/TMM.2023.3347638.

[26] V. L. Yerrabolu, "Performance Comparison of Random Forest Regressor and Support Vector Regression for Solar Energy Prediction," *Iop Conference Series Earth and Environmental Science*, vol. 1375, no. 1. 2024, doi: https://doi.org/10.1088/1755-1315/1375/1/012013.

[27] A. Primantara, "Bagging System Performance Analysis Using Artificial Neural Network, Random Forest Regression, Linear Regression, and Support Vector Regression," *Digest of Technical Papers IEEE International Conference on Consumer Electronics*. pp. 618–622, 2024, doi: https://doi.org/10.1109/ISCT62336.2024.10791247.

[28] M. Y. Shams, "Water quality prediction using machine learning models based on grid search method," *Multimed. Tools Appl.*, vol. 83, no. 12, pp. 35307–35334, 2024, doi: https://doi.org/10.1007/s11042-023-16737-4.

[29] Y. Boutahri, "Machine learning-based predictive model for thermal comfort and energy optimization in smart buildings," *Results Eng.*, vol. 22, 2024, doi: https://doi.org/10.1016/j.rineng.2024.102148.