Scientific Contributions Oil & Gas, Vol. 48. No. 4, December: 191 - 202

SCIENTIFIC CONTRIBUTIONS OIL AND GAS

Testing Center for Oil and Gas
LEMIGAS

Journal Homepage:http://www.journal.lemigas.esdm.go.id
ISSN: 2089-3361, e-ISSN: 2541-0520

Application of PCA and Machine Learning for Predicting Oil Measurement Discrepancies in Custody Transfer Systems: Understanding from an Indonesian Mature Onshore Facility

Wan Fadly[1], Fiki Hidayat[1,3], Noratikah Abu[2], Muhammad Khairul Afdhol[1,3], Dike Fitriansyah Putra[1,3], and Mulyandri[4]

[1]Department of Petroleum Engineering, Faculty of Engineering, Universitas Islam Riau
Kaharuddin Nasution Street 113, Simpang Tiga, Bukit Raya District, Pekanbaru City, Riau 28284, Indonesia

[2]Centre for Mathematical Sciences, Universiti Malaysia
Pahang 26300 Kuantan, Pahang, Malaysia

[3]Center of Energy Studies (PSE), Universitas Islam Riau
Kaharuddin Nasution Street 113, Simpang Tiga, Bukit Raya District, Pekanbaru City, Riau 28284, Indonesia

[4] PT. Pertamina Hulu Rokan
Camp Rumbai Street, Lembah Damai, Rumbai Pesisir District, Pekanbaru City, Riau 28266

Corresponding Author: Fiki Hidayat (fikihidayat@eng.uir.ac.id)

Manuscript received: October 20th, 2025; Revised: November 06th, 2025
Approved: December 08th, 2025; Available online: December 18th, 2025; Published: December 19th, 2025.

ABSTRACT - Oil measured volume discrepancies in custody transfer systems is becoming a persistent challenge, which is often caused by complex thermal, hydraulic, and compositional interactions. Therefore, this study aimed to introduce a data-driven framework incorporating Principal Component Analysis (PCA) and machine learning (ML) to identify as well as predict discrepancies at a representative onshore gathering station (GS) in Indonesia (Field-X). Major operational parameters, including gross volume, unallocated net oil, pressure, temperature, and basic sediment & Water (BS&W), were analyzed to assess the impact on volumetric imbalance. During the analysis, PCA reduced 64 correlated variables to five principal components, explaining 95% of the total variance and showing gross volume, pressure, and temperature as dominant factors. Four ML models, namely XGBoost, Random Forest, Support Vector Regression, and ElasticNet, were trained as well as validated with three-fold time series cross-validation for temporal robustness. Incorporating PCA significantly improved predictive performance, with Support Vector Regression showing the largest R^2 increase (from –0.0082 to 0.82). Results signified that discrepancies were primarily governed by thermodynamic shrinkage, temperature changes, and BS&W-related metering errors. In addition, the proposed PCA–ML framework offered an interpretable, reliable method for early detection and mitigation of oil volume discrepancies in complex production environments.

Keywords: Oil measured volume discrepancies, Time-Series Forecasting, Principal Component Analysis,

How to cite this article:

Wan Fadly, Fiki Hidayat, Noratikah Abu, Muhammad Khairul Afdhol, Dike Fitriansyah Putra, Mulyandri 2025, Application of PCA and Machine Learning for Predicting Oil Measurement Discrepancies in Custody Transfer Systems: Understanding from an Indonesian Mature Onshore Facility, Scientific Contributions Oil and Gas, 48 (4) pp. 191-202. https://doi.org/10.29017/scog.v48i4.404.

INTRODUCTION

Both onshore and offshore operations are frequently encountering oil measured volume discrepancies (OMVD) in oil and gas industry. OMVD is defined as the difference between the measured volume of crude oil received and delivered through shared transportation systems (Hermawan et al., 2021). Typical oil mixing phenomena in these systems are shown in Figure1. OMVD often occur when the total volume measured at the receiving terminal or storage tank does not match the total volume recorded at the delivery points. These inconsistencies may arise from various factors, including differences in pipeline configuration, operational pressures and temperatures, production rate fluctuations, or measurement inaccuracies in flow as well as sampling systems. The use of shared pipeline networks further causes difficulties in the issue.

When crude oil from multiple shippers is transported through a common pipeline and blended in storage facilities, determining the exact contribution and loss for each shipper becomes challenging (Badings & van Putten 2020). Consequently, the allocation process aims to fairly and accurately assign the produced as well as transported oil volumes to each shipper by correcting for parameters such as shrinkage, emulsion, and evaporation (Kanshio 2020). The challenges show the importance of developing reliable predictive systems capable of identifying anomalies in oil allocation processes. Advancements in machine learning (ML) have enabled data-driven solutions to complex prediction problems in petroleum systems. ML algorithms are capable of handling large, nonlinear, and multivariate datasets (Masini et al., 2023; Suwono & Utama 2025). In oil and gas applications, several studies have shown the potential of ML in forecasting production rates, detecting anomalies, and estimating reservoir or system parameters (Alharbi et al., 2022; Hidayat et al., 2025; Mai-Cao & Truong-Khac 2022; Rhamadhani & Saputra 2023; Song et al., 2023; Ulil et al., 2025).

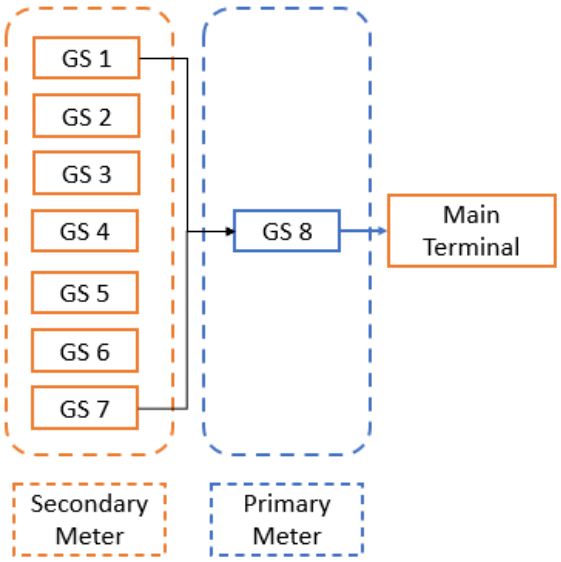

Figure 1. Mixing phenomena in shared oil transportation system.

High-dimensional operational data, such as pressure, temperature, flow rate, and basic sediment & water (BS&W) from multiple gathering station (GS), often contain redundant as well as correlated features, which reduce model generalization and lead to overfitting. To address this issue, principal component analysis (PCA) can be applied to transform correlated variables into a smaller number of uncorrelated principal components, improving model interpretability and computational efficiency (Li et al., 2021; Rangga et al., 2022; Sherif et al., 2019; Tian et al., 2024; Zhang et al., 2024). Since oil allocation data are time -dependent, model evaluation should respect the temporal structure of the data. Therefore, time series cross-validation (TSCV) is used to ensure consistent performance assessment (Bikmukhametov & Jäschke 2019; Sulandari et al., 2024; Vien et al., 2021).

This study aims to evaluate and compare the performance of several ML algorithms, such as Random Forest, XGBoost, support vector regression (SVR), Linear Regression, ElasticNet, and Bayesian Ridge in predicting OMVD values at Field-X GS. The analysis also investigates how dimensionality reduction using PCA and validation through TSCV can improve model accuracy, reduce overfitting, and improve generalization. The proposed framework offers a systematic data-driven method for predicting and mitigating OMVD, supporting operational decision-making in crude oil transportation systems.

METHODHOLGY

Dataset description

OMVD analyzed in this study for Field-X arose from the difference between the net oil volume calculated by the shippers (VΣsh, in barrels) and the volume recorded at Primary Meter (PM, in barrels). This discrepancy occurred because oil flow from each shipper was combined into a single stream before passing through Primary Meter at the main terminal.

$$Oil\ loss\ (bbl) = V_{PM} - V_{\sum sh} \quad (1)$$

$$Oil\ loss\ (\%) = \frac{V_{PM} - V_{\sum sh}}{V_{PM}} \quad (2)$$

Before preprocessing, the dataset was thoroughly assessed for quality during the analysis. No missing values or outliers were detected, ensuring a clean dataset suitable for time-series modeling without the need for additional imputation or anomaly handling procedures.

Principal component analysis

PCA was implemented as a dimensionality reduction method to address multicollinearity and improve the generalization ability of ML models. The analysis identified directions (principal components) that captured the maximum variance in the dataset and transformed correlated features into a set of linearly uncorrelated components (Salem & Hussein 2019; Vahabi & Selviah 2019). During this study, the minimum number of PCs used was generally determined by a threshold of 85% or greater of the cumulative contribution rate (Han & Kwon 2021). Data collection was conducted from June 6 to August 15, 2023, during this study. Observations at each GS provided several major data points, including BS&W, fluid pressure and temperature in the pipeline, unallocated net volume (gross volume with blending factor correction), allocated volume (gross volume from the main terminal), metering factor data, as well as gross volume data.

PCA–ML incorporation process. Operational variables collected from multiple GS (GS1–GS7) were first passed through PCA, where the features were transformed into a reduced set of uncorrelated principal components (PC1, PC2, …, n-PC). These components then served as inputs to ML algorithms used for OMVD prediction.

The system used a control parameter that customized the selection process for each original tree, improving its flexibility in adapting to specific data sets rather than to other sets of trees (Chen & Guestrin 2016; Wood 2023).

Model construction

Several ML models were used to identify which model performed best in predicting OMVD values. This section reviewed the specific characteristics and algorithmic configurations of each model, as shown in Table 1.

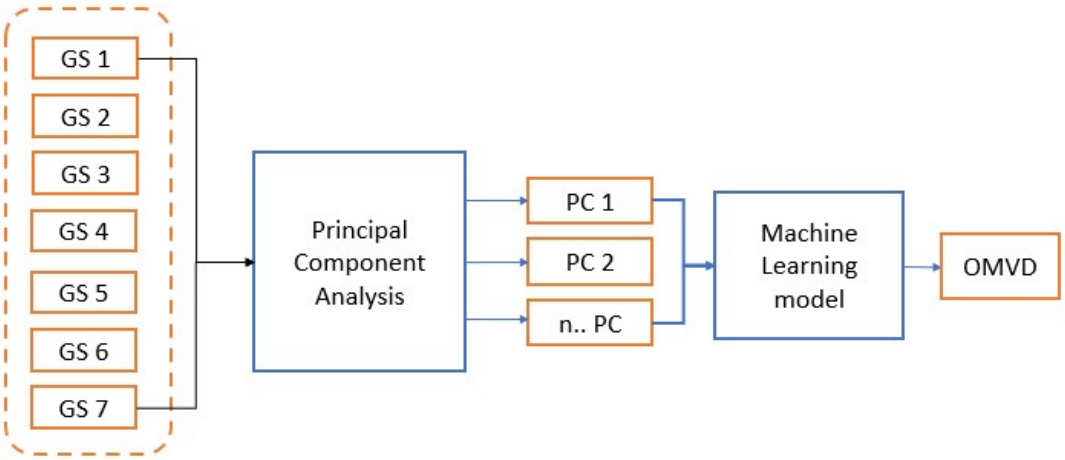

Figure 2. PCA–ML workflow for OMVD prediction.

Table 1. Review of ML models used.

Model	Description	Algorithm Parameters
XGB	The system used a control parameter that customized the selection process for each original tree, improving its flexibility in adapting to specific data sets rather than to other sets of trees (Chen & Guestrin, 2016; Wood, 2023).	• The number of estimators was 1000 • Max depth was 6 • Learning rate was 0.1 • Alpha is 0.8
RF	An ensemble tree-based model using bootstrap aggregation (bagging) for improved variance reduction and interpretability (Ilic et al., 2021; Nemer, 2024).	• Number of estimators was 1000 • Max depth was 10 • Min sample leaf was 3
SVR	A kernel-based regression model that discover the optimal hyperplane minimizing prediction error (Dsouza, 2024; Pisner & Schnyer, 2020; Wardhana et al., 2021).	• Kernel type was RBF • SVR cost was 1 • Gamma was 0.0001 • Epsilon was 0.005
MLR	A linear model assuming a direct relationship between predictors and the target variable (Alharbi et al., 2022).	None
EN	A hybrid regularization model combining Lasso (L1) and Ridge (L2) penalties to improve model stability and feature selection (Al-Jawarneh et al., 2022).	• Alpha was 0.01 • L1 rasio was 0.1 • Maxium iteration was 500
BR	A probabilistic linear regression model that estimated weight distributions under Gaussian priors (Effrosynidis et al., 2023).	• Alpha 1 was 0.1 • Alpha 2 was 0.1 • Lambda was 0.001 • Cost was 0.000001

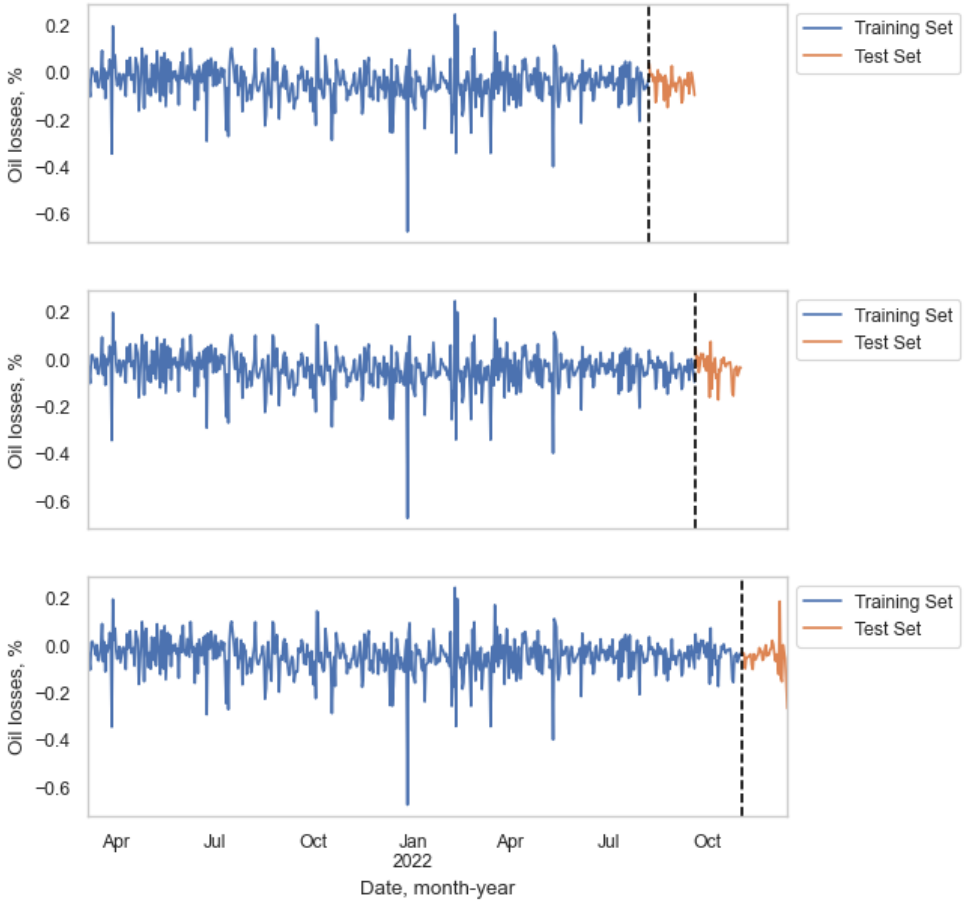

Figure 3. Schematic representation of validation scheme.

Model validation and evaluation

TSCV was used to ensure model robustness in temporal prediction, rather than conventional random k-fold validation. Traditional cross-validation assumed independent and identically distributed samples, which was not suitable for time-dependent data (Botache et al., 2023).

TSCV method divided the dataset sequentially, allowing the training data to often precede the validation data in time (Bikmukhametov & Jäschke 2019). In this study, the dataset was divided into 80% training and 20% testing, with three folds applied for model validation, as shown in Figure 3. The method prevented data leakage and provided a realistic assessment of predictive performance on unobserved future data.

Model evaluation was conducted using three standard forecasting performance metrics, namely mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). These metrics were selected to measure both absolute deviation and predictive goodness-of-fit.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y^i - \hat{y}^i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^i - \hat{y}^i)^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y^i - \hat{y}^i)^2}{\sum_{i=1}^{n}(\bar{y}^i - \hat{y}^i)^2} \quad (5)$$

where n was the number of test data, y^i, $\hat{y}^i$ and $\bar{y}^i$ signified the actual value, predicted value, and average of discrepancy values, respectively.

RESULT AND DISCUSSION

Result

This study evaluated several ML algorithms to predict OMVD in Field-X GS network. The application of PCA provided a more compact yet highly informative representation of the operational dataset. Concerning the original 64 correlated variables, the first five principal components captured approximately 95% of the total variance. Furthermore, loading values were examined to identify the most influential variables in each principal component (Parhizkar et al., 2021).

Examination of the loading values showed that major operational parameters, particularly gross volume, unallocated net oil, average pressure, and average temperature, dominated the principal component structure. This signified that volumetric fluctuations and thermohydraulic conditions across GS were the primary drivers influencing OMVD behavior.

PCA outcomes were shown in Figure 3 and Table 2. PC1 and PC2 were primarily associated with gross volume and unallocated net oil from multiple GS, reflecting the variability in transported volume as well as inter-station imbalances. Consequently, PC3 through PC5 were characterized by pressure and temperature variables, capturing the physical dynamics of fluid conditions along the pipeline system. These patterns confirmed that PCA reduced dimensionality and also signified latent structures, physically following the mechanisms behind measured volume discrepancies, and served as compact yet informative inputs for ML models. Model evaluation using three-fold TSCV without PCA was shown in Table 3 and 4. The analysis also showed the time series plot for the model with and without PCA in Figure 4.

After applying PCA, model stability and accuracy improved significantly during the analysis. SVR model showed the greatest improvement, with R^2 rising from -0.0082 to 0.82 and RMSE reduced by 62%. Ensemble models also improved, with XGBoost achieving R^2 0.86 and Random Forest R^2 0.84, while maintaining error consistency across folds. Linear models showed smaller and consistent improvements, with RMSE reductions between 8-15%. Across all models, PCA decreased overfitting and improved cross-validation reproducibility.

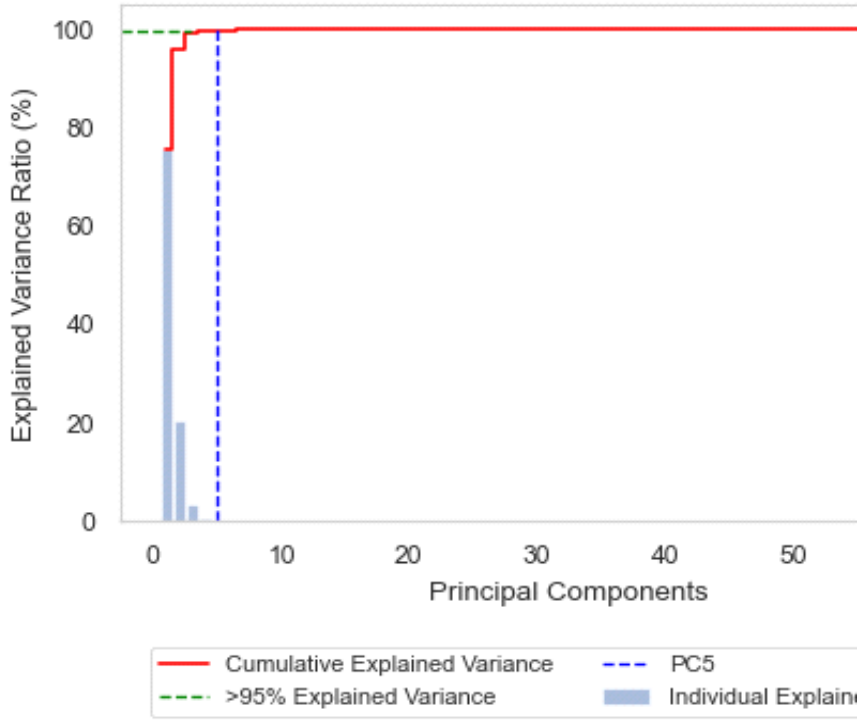

Figure 4. Cumulative explained variance Vs. number of principal components.

Table 2. Most influential variables in each principal component based on PCA loading values.

Principal Components	Key Variable
PC 1	• **Allocated GS 5** • **Gross Volume GS 5** • **Unallocated Net GS 5**
PC 2	• **Allocated GS 8** • **Gross Volume GS 8** • **Unallocated Net GS 8**
PC 3	• **Gross Volume GS 7** • **Unallocated Net GS 7** • **Gross Volume GS 4**
PC 4	• **Average Pressure GS 7** • **Average Pressure GS 4** • **Average Pressure GS 2** • **Average Pressure GS 3**
PC 5	• **Gross Volume GS 6** • **Average Temperature GS 6** • **Average Temperature GS 7** • **Average Temperature GS 3**

Table 3. Error metrics from each model without PCA.

Model	Error metric	Fold			Average
		1th	2th	3th	
XGB	MAE	0.0156	0.444	0.040	0.033
	RMSE	0.0185	0.053	0.117	0.063
	R^2	0.81	-0.09	-2.5	-0.60
RF	MAE	0.009	0.05	0.03	0.034
	RMSE	0.012	0.063	0.087	0.054
	R^2	0.92	-0.53	-0.98	-0.19
MLR	MAE	0.009	0.058	0.025	0.031
	RMSE	0.011	0.082	0.037	0.043
	R^2	0.92	-1.5	0.6	-0.007
SVR	MAE	0.032	0.038	0.037	0.035
	RMSE	0.042	0.051	0.062	0.052
	R^2	-0.008	-0.014	-0.002	-0.008
BR	MAE	0.010	0.021	0.017	0.016
	RMSE	0.012	0.028	0.029	0.023
	R^2	0.9	0.69	0.77	0.79
EN	MAE	0.009	0.022	0.017	0.016
	RMSE	0.012	0.028	0.027	0.022
	R^2	0.91	0.69	0.8	0.8

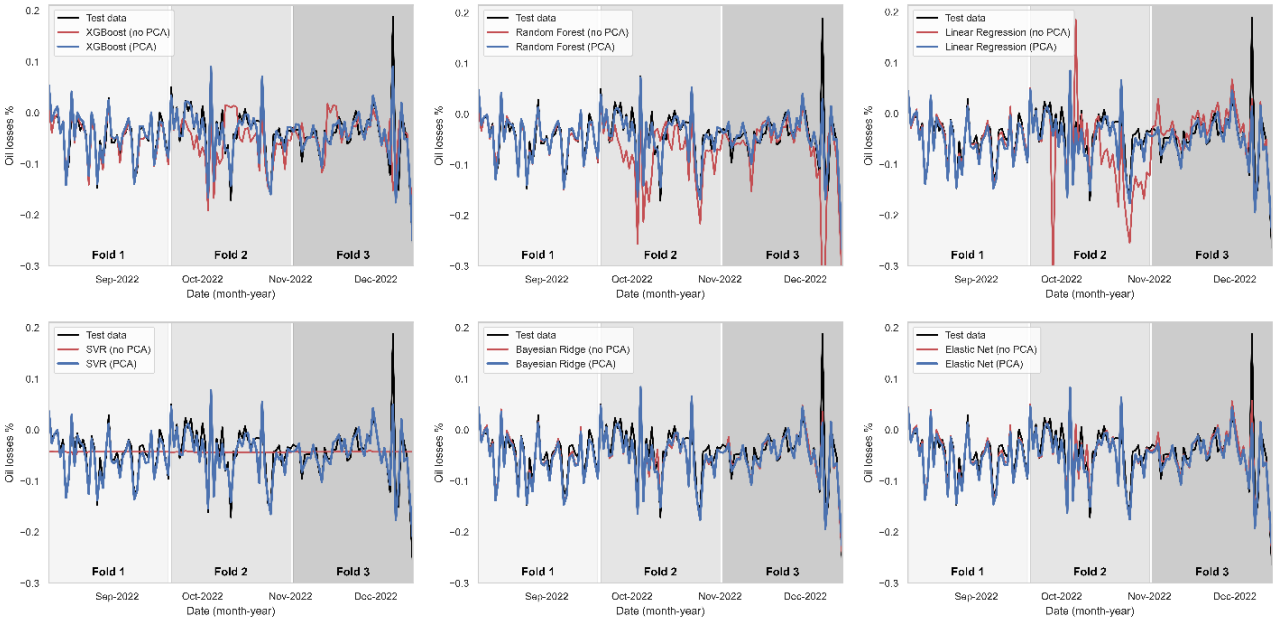

Figure 5. Model performance comparison

Discussion

The incorporated analysis of model results and principal component patterns provided a deeper understanding of the physical drivers of OMVD and the behavior of ML algorithms under different data structures. OMVD in petroleum pipeline systems originated from a combination of thermal, hydraulic, and compositional effects that altered the measured volume during transportation (Emeke 2019; Vakilinejad et al., 2017).

The dominance of gross volume, unallocated net oil, temperature, and pressure variables in the first five principal components signified that these operational conditions were the main contributors to OMVD. As crude oil traveled from multiple GS to the central metering point, pressure reduction along the flowline led to gas liberation and volumetric shrinkage, while temperature fluctuation promoted phase instability as well as emulsion formation. The presence of BS&W compounded this effect by introducing emulsified layers that distorted metering accuracy, further widening the observed difference between sent and received volumes (Nengkoda 2011).

PCA–ML framework showed potential as a practical diagnostic tool for OMVD monitoring from an operational perspective. The identification of pressure and temperature as dominant contributors reinforced the importance of maintaining stable flowline conditions through insulation, backpressure regulation, and continuous temperature compensation (Hermawan et al., 2021). The contribution of unallocated net oil also showed the need for better reconciliation and calibration among GS to minimize metering bias (Badings & van Putten 2020; Kanshio 2020).

Moreover, PCA-derived components served as early indicators of anomalies, which included sensor drift, blending mismatch, or pipeline imbalance, supporting proactive maintenance and allocation transparency. Some models were unable to learn the pattern of the actual oil loss value without PCA, which occurred when the amount of training data was less. High fluctuation in the dataset caused an overfitting effect on the performance of the models (Bikmukhametov & Jäschke 2019; Nugroho & Husin 2022; Rhamadhani & Saputra 2023; Song et al., 2023).

Ensemble models such as XGBoost and Random Forest showed the strongest predictive stability when trained on PCA-transformed data in terms of algorithmic behavior. The reduced dimensionality helped the models focus on representative features (Zhang et al., 2024), improving the ability to generalize across

Table 4: Error metrics from each model using PCA.

Model	Error metric	Fold			Average
		1th	2th	3th	
XGB	MAE	0.009	0.015	0.014	0.013
	RMSE	0.012	0.022	0.022	0.019
	R^2	0.91	0.8	0.86	0.86
RF	MAE	0.009	0.014	0.015	0.013
	RMSE	0.011	0.020	0.029	0.02
	R^2	0.92	0.83	0.77	0.84
MLR	MAE	0.013	0.020	0.017	0.017
	RMSE	0.015	0.027	0.032	0.025
	R^2	0.86	0.71	0.72	0.77
SVR	MAE	0.010	0.018	0.016	0.015
	RMSE	0.012	0.024	0.027	0.021
	R^2	0.9	0.76	0.8	0.82
BR	MAE	0.013	0.020	0.017	0.017
	RMSE	0.015	0.027	0.032	0.025
	R^2	0.86	0.71	0.72	0.77
EN	MAE	0.012	0.020	0.017	0.016
	RMSE	0.015	0.027	0.032	0.024
	R^2	0.86	0.72	0.72	0.77

operational periods without overfitting to transient anomalies. SVR model benefited the most from PCA since the kernel-based method relied heavily on orthogonal feature spaces by eliminating correlated variance (Li et al., 2021; Osah & Howell 2023). PCA allowed SVR to better capture the nonlinear interplay between thermal and volumetric parameters that governed OMVD. However, regularized linear models such as ElasticNet and Bayesian Ridge showed that excessive dimensional compression caused over-regularization (Naufal & Metra 2021; OKON et al., 2024; Sola-Aremu 2019). The results showed that dimensionality reduction clarified the fundamental structure of oil transport data, improved model interpretability, and supported more reliable forecasting of OMVD trends. Therefore, PCA–ML framework represented a physically grounded, data-driven method for managing OMVD in shared pipeline networks, providing both predictive accuracy and operational understanding for improved production accountability.

CONCLUSION

In conclusion, the application of PCA combined with TSCV effectively improved the predictive performance and stability of all tested models. SVR model showed the greatest improvement, with its R^2 value increasing from –0.0082 to 0.82, while ensemble models such as XGBoost and Random Forest achieved accuracies of more than 0.88 under temporal validation. PCA successfully reduced 64 correlated variables into five principal components that captured approximately 95% of the total data variance, dominated by gross volume, unallocated net oil, pressure, and temperature. These results showed that OMVD were a physically driven phenomenon governed by pressure decline, temperature fluctuation, and BS&W variation rather than random measurement noise. The incorporation of PCA and TSCV enabled models to generalize more effectively across time-dependent operational data, while improving interpretability by isolating major thermohydraulic relationships and filtering noise. The developed PCA–ML framework provided a reliable and explainable tool for OMVD diagnosis, supporting early detection as well as mitigation of discrepancies in multi-station gathering systems.

ACKNOWLEDGEMENT

This study was conducted with the support of the Matching Grant UMP–UIR 2024 provided by Universitas Islam Riau (UIR) in collaboration with Universiti Malaysia Pahang (UMP), under Contract No. 1235/KONTRAK/P-P-MGUMP/ DPPM-UIR/10-2024.

GLOSSARY OF TERMS

Symbol	Definition	Unit
RF	Random Forest	
XGB	Extreme Gradient Boosting	
OMVD	Oil Measured Volume Discrepancy	[barrel]
TSCV	Time Series Cross-Validation	
SVR	Support Vector Regression	
GS	Gathering Station	
BR	Bayesian Ridge	
EN	Elastic Net	
ML	Machine Learning	
PC	Principal Component	
MLR	Multiple Linear Regression	
BS&W	Basic Sediment and Water	[%]
TSCV	Time Series Cross-Validation	
PCA	Principal Component Analysis	

REFERENCES

Alharbi, R., Alageel, N., Alsayil, M., & Alharbi, R. (2022). Prediction of oil production through linear regression model and big data tools. International Journal of Advanced Computer Science and Applications, 13(12).

Al-Jawarneh, A. S., Ismail, M. T., Awajan, A. M., & Alsayed, A. R. M. (2022). Improving accuracy models using elastic net regression method based on empirical mode decomposition. Communications in Statistics-Simulation and Computation, 51(7), 4006–4025.

Badings, T. S., & van Putten, D. S. (2020). Data validation and reconciliation for error correction and gross error detection in multiphase allocation systems. Journal of Petroleum Science and Engineering, 195, 107567.

Bikmukhametov, T., & Jäschke, J. (2019). Oil production monitoring using gradient boosting machine learning algorithm. Ifac-Papersonline, 52(1), 514–519.

Botache, D., Dingel, K., Huhnstock, R., Ehresmann, A., & Sick, B. (2023). Unraveling the Complexity of Splitting Sequential Data: Tackling Challenges in Video and Time Series Analysis. ArXiv Preprint ArXiv:2307.14294.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785 –794.

Dsouza, N. A. (2024). Evaluation of Machine Learning Algorithms for Flow Rate Estimation in Oil and Gas Industry [Master's thesis, University of South-Eastern Norway]. www.usn.no

Effrosynidis, D., Spiliotis, E., Sylaios, G., & Arampatzis, A. (2023). Time series and regression methods for univariate environmental forecasting: An empirical evaluation. Science of The Total Environment, 875, 162580.

Emeke, K. B. C. (2019). A novel model developed for forecasting oilfield production using multivariate linear regression method. Journal of Science and Technology Study, 29(2), 579–591.

Han, D., & Kwon, S. (2021). Application of machine learning method of data-driven deep learning model to predict well production rate in the shale gas reservoirs. Energies, 14(12), 3629.

Hidayat, F., Nasution, A. H., Ambia, F., & Putra, D. F. (2025). Leveraging Large Language Models for Discrepancy Value Prediction in

Custody Transfer Systems: A Comparative Analysis of Probabilistic and Point Forecasting Methods. IEEE Access.

Ilic, I., Görgülü, B., Cevik, M., & Baydoğan, M. G. (2021). Explainable boosted linear regression for time series forecasting. Pattern Recognition, 120, 108144.

Kanshio, S. (2020). A review of hydrocarbon allocation methods in the upstream oil and gas industry. Journal of Petroleum Science and Engineering, 184, 106590.

Li, X., Zhang, L., Khan, F., & Han, Z. (2021). A data-driven corrosion prediction model to support digitization of subsea operations. Process Safety and Environmental Protection, 153, 413–421.

Mai-Cao, L., & Truong-Khac, H. (2022). A comparative study on different machine learning algorithms for petroleum production forecasting. Improved Oil and Gas Recovery, 6.

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. Journal of Economic Surveys, 37(1), 76–111.

Naufal, A. A., & Metra, S. (2021). A digital oilfield comprehensive study: Automated intelligent production network optimization. SPE Asia Pacific Oil and Gas Conference and Exhibition, D031S026R003.

Nemer, Z. N. (2024). Oil and Gas Production Forecasting Using Decision Trees, Random Forst, and XGBoost. Journal of Al-Qadisiyah for Computer Science and Mathematics, 16(1), 9–20.

Nengkoda, A. (2011). The role of crude oil shrinkage in heavy mix light crude in main oil pipeline: case study Oman. SPE International Heavy Oil Conference and Exhibition, SPE-148925.

Nugroho, A., & Husin, A. (2022). Analisis Performa Random Forest Menggunakan Normalisasi Atribut. SISTEMASI: Jurnal Sistem Informasi, 11(1), 186–196.

Okon, J., Udoh, T., & Emenka, B. (2024). Prediction of Interfacial Tension Using Machine Learning: A Review of Applied Techniques in Petrochemical/Reservoir Engineering.

Osah, U., & Howell, J. (2023). Predicting oil field performance using machine learning programming: a comparative case study from the UK continental shelf. Petroleum Geoscience, 29(1), petgeo2022-071.

Parhizkar, T., Rafieipour, E., & Parhizkar, A. (2021). Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction. Journal of Cleaner Production, 279, 123866.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101–121). Elsevier.

Rangga, A., Widyasari, Y. D. L., & Sahid, D. S. S. (2022). Integrated production facilities clustering and time-series forecasting derived from large dataset of multiple hydrocarbon flow measurement. Science, Technology and Communication Journal, 2(2), 32–45.

Rhamadhani, D. A., & Saputra, E. E. D. (2023). Analisa Model Machine Learning dalam Memprediksi Laju Produksi Sumur Migas 15/9-F-14H. Journal of Sustainable Energy Development, 1(1), 48–55.

Salem, N., & Hussein, S. (2019). Data dimensional reduction and principal components analysis. Procedia Computer Science, 163, 292–299.

Sherif, S., Adenike, O., Obehi, E., Funso, A., & Eyituoyo, B. (2019). Predictive data analytics for effective electric submersible pump management. SPE Nigeria Annual International Conference and Exhibition, D033S019R003.

Sola-Aremu, O. (2019). An inferable machine learning method to predicting PVT properties of Niger delta crude oil using compositional data. SPE Annual Technical Conference and Exhibition?, D023S103R021.

Song, L., Wang, C., Lu, C., Yang, S., Tan, C., & Zhang, X. (2023). Machine Learning Model of Oilfield Productivity Prediction and Performance Evaluation. Journal of Physics: Conference Series, 2468(1), 012084.

Sulandari, W., Yudhanto, Y., Subanti, S., Zukhronah, E., & Subarkah, M. Z. (2024).

Implementing Time Series Cross Validation to Evaluate the Forecasting Model Performance. KnE Life Sciences, 229–238.

Suwono, S., & Utama, D. N. (2025). Estimation of Well Flowing Bottomhole Pressure (FBHP) Using Machine Learning. Scientific Contributions Oil and Gas, 48(3), 33–44. https://doi.org/10.29017/scog.v48i3.1851.

Tian, F., Fu, Y., Liu, X., Li, D., Jia, Y., Shao, L., Yang, L., Zhao, Y., Zhao, T., & Yin, Q. (2024). A Comprehensive Evaluation of Shale Oil Reservoir Quality. Processes, 12(3), 472. https://doi.org/10.29017/scog.v48i3.1851.

Ulil, M. R., Winardhi, S., & Dinanto, E. (2025). Machine Learning-Based Prediction of Shear Wave Velocity: Performance Evaluation of Bi-Gru, Ann, and The Greenberg-Castagna Empirical Method. Scientific Contributions Oil and Gas, 48(3), 133–144. https://doi.org/10.29017/scog.v48i3.1797.

Vahabi, N., & Selviah, D. R. (2019). Dimensionality reduction and pattern recognition of flow regime using acoustic data. Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2, 880–891.

Vakilinejad, A., Ahmad, W., & Vakili-Nezhaad, G. (2017). Volumetric Behavior Study of Petroleum Fluids Mixtures Through Shrinkage Factor. ICTEA: International Conference on Thermal Engineering, 2017.

Vien, B. S., Wong, L., Kuen, T., Rose, L. F., & Chiu, W. K. (2021). A machine learning method for anaerobic reactor performance prediction using long short-term memory recurrent neural network. Struct. Health Monit, 18, 61.

Wardhana, S. G., Pakpahan, H. J., Simarmata, K., Pranowo, W., & Purba, H. (2021). Algoritma komputasi machine learning untuk aplikasi prediksi nilai total organic carbon (TOC). Lembaran Publikasi Minyak Dan Gas Bumi, 55(2), 75–87. https://doi.org/10.29017/LPMGB.55.2.606.

Wood, D. A. (2023). Geomechanical brittleness index prediction for the Marcellus shale exploiting well-log attributes. Results in Engineering, 17, 100846.

Zhang, Y., Zhang, G., Zhao, W., Zhou, J., Li, K., & Cheng, Z. (2024). Total organic carbon content estimation for mixed shale using Xgboost method and implication for shale oil exploration. Scientific Reports, 14(1), 20860.