Explainable Machine Learning Framework for Predicting Retention Time Shifts in Biodiesel Gas Chromatography

Montassar T.bouzidi[a], **Nur Alif Fathurrahman**[b], **Listya Eka Anggraini**[b],**Abdulaziz Al-Saadi**[a]

[a] *Department of Chemistry, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*

chem@kfupm.edu.sa

[b] *New & Renewable Energy (NRE)-Biofuel Laboratory, Department of Product Application Technology, Testing Centre for Oil and Gas LEMIGAS, South Jakarta 12230, Indonesia info.lemigas@esdm.go.id*

Abstract: Retention time shifts in gas chromatography (GC) complicate reliable analysis and quality control in biodiesel production. These variations, influenced by factors such as instrument drift, column aging, and changes in operational parameters, require accurate predictive methods. We propose an explainable machine learning (ML) framework using linear regression combined with SHapley Additive exPlanations (SHAP) to predict and interpret retention time shifts. The model, trained on experimental biodiesel GC data, achieved a high predictive performance (MAE = 0.114, MSE = 0.020, R^2 = 0.992). SHAP analysis highlighted key influencing factors, including temperature rates, hold time, and column age positively impacting shifts, whereas flow rates and laboratory temperature reduced variability. This transparent ML approach enhances prediction accuracy, interpretability, and reliability, making it highly suitable for integration into biodiesel quality assurance workflows.

1. Introduction: Retention time variability in gas chromatography (GC) negatively impacts biodiesel quality control by complicating chromatographic data interpretation. These shifts result from operational variations, instrument drift, and column aging. Existing methods, including manual corrections and polynomial regression, are limited in accuracy and scalability. Although recent machine learning approaches provide improved accuracy, many remain "black boxes" lacking transparency, hindering their adoption. This study introduces an interpretable linear regression model coupled with SHAP to predict and clearly interpret retention shifts, offering significant advantages in analytical reliability, transparency, and practical applicability.

2. Related work: Traditionally, GC retention time shifts have been managed through manual methods, which, while simple, are labor-intensive and prone to errors (Hinshaw, 2018). Polynomial regression techniques have automated this process, but their performance diminishes with data complexity (Zhao et al., 2013). Recent ML applications, like Random Forest and neural networks, have provided improved accuracy but suffer from limited interpretability (Zang et al., 2023). Explainable ML using SHAP has improved interpretability in fields such as pharmaceutical GC analysis, yet its specific use in biodiesel analysis remains rare (Singh et al., 2023). Our approach leverages linear regression combined with SHAP, addressing these limitations by offering both accuracy and interpretability specifically tailored for biodiesel analysis.

3. Methodology

3.1 Data collection and Description

Experimental chromatographic data were obtained from Balai Besar Pengujian Minyak dan Gas Bumi (LEMIGAS), located at Jalan Ciledug Raya, Kav. 109 Cipulir, Kebayoran Lama, 12230, Indonesia. The dataset consisted of 500 biodiesel samples, specifically targeting glycerol analysis. Detailed parameters and features collected are summarized clearly in Table 1.

Table 1: Biodiesel GC experimental dataset parameters

Parameters		
Sample	Biodiesel	
Compound of Interest	Glycerol	
No. of data	500 datasets	
Features	Flow Rate (mL/min)	Detector Temp (°C)
	Initial Temp (°C)	Column Age (months)
	Rate 1 Temp (°C)	Lab Temp (°C)
	Rate 2 Temp (°C)	Humidity (%)
	Rate 3 Temp (°C)	Retention Time (min)
	Hold Time (min)	Peak Shifting (min)

3.2 Model Development and Validation

Data Preprocessing: Checked for outliers and missing values, then normalized features to improve model stability.

Linear Regression Model: which is Developed using Python's Scikit-learn library, and Trained with 80% of the dataset (400 samples), validated with the remaining 20%.

Interpretability using SHAP: SHAP values calculated clearly illustrate feature contributions to retention time shifts.

4. Results and Discussion

4.1 Model Performance:

The performance of the linear regression model was evaluated clearly using training and testing datasets (Table 2).

Explainable Machine Learning Framework for Predicting Retention Time Shifts in Biodiesel Gas Chromatography

Montassar T.bouzidi[a], **Nur Alif Fathurrahman**[b], **Listya Eka Anggraini**[b],**Abdulaziz Al-Saadi**[a]

[a] *Department of Chemistry, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*

chem@kfupm.edu.sa

[b] *New & Renewable Energy (NRE)-Biofuel Laboratory, Department of Product Application Technology, Testing Centre for Oil and Gas LEMIGAS, South Jakarta 12230, Indonesia info.lemigas@esdm.go.id*

Table 2: Linear regression model performance metrics

Evaluation matrics	MAE	MSE	R^2
Training Performance:	0.116	0.021	0.989
Testing Performance:	0.114	0.020	0.992

The model demonstrated high accuracy on both datasets, with minimal difference indicating strong generalization and robustness. Figure 1 clearly illustrates the actual versus predicted retention time shifts, demonstrating the model's high predictive capability.

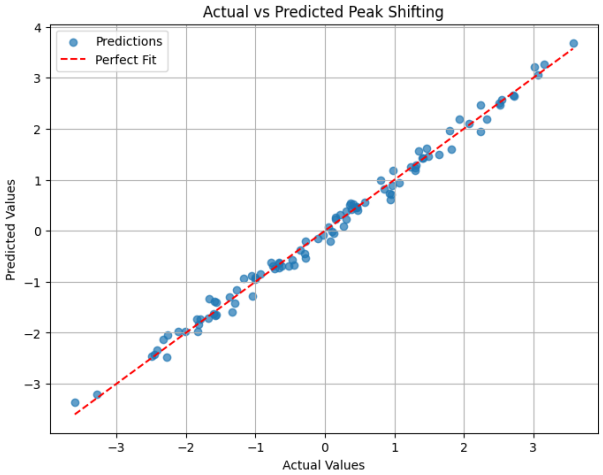

Figure 1: Actual vs. Predicted Peak Shifting

4.2 Feature Importance and Interpretation

Feature importance analysis was conducted using SHAP values, clearly illustrating the impact of each operational parameter (Figure 2).

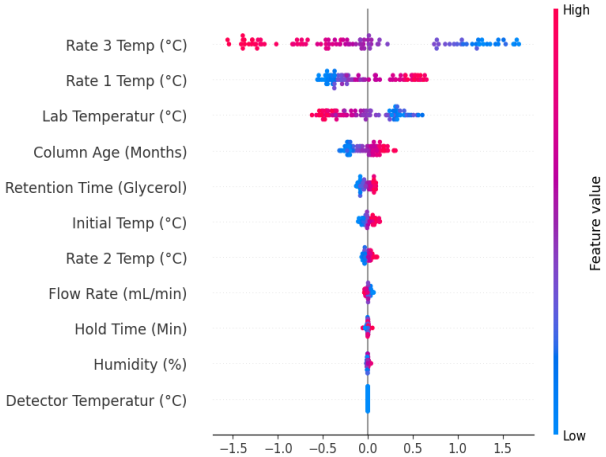

Figure 2: SHAP Analysis of Feature Importance

Positive Impact (increase peak shifts): Rate 2 temperature, Rate 1 temperature, Hold time, Column age, Initial temperature, Retention time of glycerol.
Negative Impact (reduce peak shifts)**:** Rate 3 temperature, Flow rate, Laboratory temperature.
Minimal or No Impact: Humidity and Detector temperature.
To further quantify these influences, coefficients from the linear regression model are provided in Table 3.

Table 3: Linear regression model coefficients

Feature	Coefficient
Rate 2 Temp (°C)	0.901
Rate 1 Temp (°C)	0.753
Hold Time (min)	0.626
Column Age (months)	0.255
Initial Temp (°C)	0.246
Retention Time (Glycerol)	0.187
Humidity (%)	0.010
Detector Temp (°C)	~0.000
Lab Temp (°C)	-0.327
Flow Rate (mL/min)	-0.728
Rate 3 Temp (°C	-0.875

The clear interpretation of feature importance provides valuable insights for targeted operational adjustments, optimizing biodiesel GC performance and improving quality control consistency.

5.Conclusion:

This study presents a novel and practical explainable machine learning framework designed to accurately predict and clearly interpret retention time shifts in biodiesel gas chromatography. By integrating linear regression with SHAP analysis, the approach not only achieves outstanding predictive accuracy but also provides transparent insights into key operational factors, enabling targeted and efficient process optimization. This clear interpretability directly supports improved decision-making, reduces operational uncertainty, and boosts reliability in biodiesel production. The adaptability of this framework further expands its potential applications into diverse analytical sectors, including pharmaceuticals, petrochemicals, and environmental monitoring, substantially benefiting overall production efficiency and quality assurance processes.

Explainable Machine Learning Framework for Predicting Retention Time Shifts in Biodiesel Gas Chromatography

Montassar T.bouzidi[a], **Nur Alif Fathurrahman**[b], **Listya Eka Anggraini**[b],**Abdulaziz Al-Saadi**[a]

[a] *Department of Chemistry, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*

chem@kfupm.edu.sa

[b] *New & Renewable Energy (NRE)-Biofuel Laboratory, Department of Product Application Technology, Testing Centre for Oil and Gas LEMIGAS, South Jakarta 12230, Indonesia info.lemigas@esdm.go.id*

References:

[1] J.V. Hinshaw. Troubleshooting GC Retention-Time, Efficiency, and Peak-Shape Problems. *Chromatography Online*, 2018. https://www.chromatographyonline.com/view/troubleshooting-gc-retention-time-efficiency-and-peak-shape-problems. (Accessed: 13.03.2025).

[2] L. Zhao, L. Zhang, and G. Xu. Difference Equation Model for Isothermal Gas Chromatography Retention. *PLoS ONE*, 8(2): e56219, 2013. https://doi.org/10.1371/journal.pone.0056219. (Accessed: 13.03.2025).

[3] W. Zang, W. Jia, H. Lu, and H. Zhai. Retention Time Trajectory Matching for Target Compound Peak Identification in Chromatographic Analysis. *Molecules*, 28(11):4573, 2023. https://doi.org/10.3390/molecules28114573. (Accessed: 13.03.2025).

[4] A. Singh, N. Thakur, and V. Sharma. Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Models. *Journal of Biomedical Informatics*, 138:104340, 2023. https://doi.org/10.1016/j.jbi.2023.104340. (Accessed: 13.03.2025).

Acknowledgments:

The authors thank the Chemistry Department for their valuable assistance and for providing access to the software tools necessary for conducting this research. Special thanks are extended to the Bali Biodiesel Research Center for generously supplying experimental data and facilitating validation procedures, which significantly contributed to the successful completion of this study.