Efficient and Simplified Modeling for Kerosene Processing Quality Detection Using Partial Least Squares-Discriminant Analysis Regression

Hayder M. Issa* and Rezan H. Hama Salih

Department of Chemistry, University of Garmian, Kalar, Sulaymaniyah Province, 46021, Kurdistan Region – F.R. Iraq

***Abstract*—Kerosene from various refineries and crudes is used for heating and other purposes in many countries like Iraq; therefore, it is important to identify its source to recognize and tax any adulteration. In this study, a fast classification technique for kerosene marketed in Iraq was developed with the goal of identifying its quality. The samples were categorized using a supervised partial least squares discriminant analysis (PLS-DA) approach. Multivariate analyses using agglomerative hierarchal clustering and principal component analysis were utilized to identify outliers and sample dissimilarities. The dataset was divided into calibration and prediction sets. The prediction set was used to evaluate the model's separation performance. The Q^2 cross-validation was applied. The PLS-DA models achieved significant accuracy, sensitivity, and specificity, showing strong segregation ability, notably for the calibration set (100% accuracy and 1.00 sensitivity). It was found that kerosene processing can be classified rapidly and non-destructively without the need for complicated analyses, demonstrating the best results for classification even when compared with the classification outcomes of other fuels. This PLS-DA approach has never been looked at before for process quality detection, and the results are comparable to direct kerosene classification with soft independent modeling of class analogy and support vector machines.**

***Index Terms*—Supervised discrimination technique, Modeling, Kerosene processing, Quality control, PLS-DA, Machine learning tool.**

I. Introduction

Kerosene is a mixture of hydrocarbons with carbon atom counts ranging from 6 to 16, and kerosene can be utilized in a variety of applications, including as an aviation fuel and a home heating fuel. It is isolated as a straight-chain component of petroleum; kerosene is produced primarily through fractional distillation of crude oil (Lam, et al., 2012). N-alkane, alkyl benzene, and naphthalene are the main components of kerosene (Kaltschmitt and Deutschmann, 2012). Its volatility is between that of gasoline and diesel fuel, and its boiling point range is between 150 and 350°C.

The refineries at Bazian and Kirkuk are the main sources of kerosene for the region of study in north-eastern Iraq. A total of 34,000 barrels/day is produced at the 2009-founded Bazian refinery, which uses only atmospheric distillation and hydrotreating equipment. The Taq-Taq oil field is the principal source of crude oil for the Bazian refinery (Ali and Khodakarami, 2015). The Kirkuk oil field supplies the crude for the Kirkuk refinery, which began operations in 1938 and has seen its capacity rise over the years to the current 30,000 barrels/day. Atmospheric distillation, vacuum distillation, catalytic reforming, and hydrotreating are just some of the process units in the Kirkuk refinery (Abdullah and Daij, 2021). Both Taq-Taq and Kirkuk crudes can be refined into kerosene at about the same rate (25.6% for Taq-Taq and 23.6% for Kirkuk), but Taq-Taq crude has a higher API (47.52) and is more expensive (Karim, Khanaqa and Shukur, 2017, Naman, et al., 2019). When comparing the Bazian refinery to the Kirkuk refinery, the Nelson complexity index, which is used to define the sophistication of a refinery (Kaiser, 2017), for the former is 2.26 whereas the latter is 2.03 (Abdullah and Daij, 2021).

Therefore, there must be a method to categorize and identify the source of locally produced kerosene because each type has a distinct composition, which affects performance as well as the possibility of adulteration and taxation. Several standards can be used to control the quality of middle distillates like kerosene. They are time-consuming, require many samples, and have expensive measurement equipment. It has proven possible to identify sources, detect adulteration, or classify fuels using chemometrics analysis with spectral or property input data (Barra, et al., 2020, De Paulo, et al., 2014, Comesaña-García, et al., 2013, Dago Morales, et al., 2008). The principle component analysis (PCA) and hierarchical cluster analysis (HCA) multivariate calibration techniques have been used by Tanaka, et al. (2011) to detect solvent traces in gasoline. They found 83.8% and 77.1% sensitivity results for calibration and external sets, respectively, by

ARO-The Scientific Journal of Koya University
Vol. XII, No. 1 (2024), Article ID: ARO.11515. 8 pages
Doi: 10.14500/aro.11515
Received: 08 January 2024; Accepted: 21 March 2024
Regular research paper: Published: 10 May 2024
Corresponding author's e-mail: hayder.mohammed@garmian.edu.krd

adopting the soft independent modeling of class analogy (SIMCA) algorithm. Dadson, Pandam, and Asiedu (2018) have looked at the possibility of classifying gasoline based on four added adulterants using the same approach except for HCA. The SIMCA classification model has a sensitivity of 100% for the calibration set and 75% for the external set.

Furthermore, Mohammadi, et al. (2020) employed attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy with partial least squares discriminant analysis (PLS-DA) to sort crude oil samples into groups. This created a model that was 100% accurate in terms of both sensitivity and specificity. Mazivila, et al. (2015) used the mid-infrared spectroscopy (MIR) data to sort biodiesel samples into groups based on their types and routes. They did this using a PLS-DA-based classification algorithm. The PLS-DA classification model yielded identical accuracy values of 100%.

Kerosene is relatively understudied compared to gasoline and biodiesel. Innovatively, this work categorizes kerosene by source conformance using a multivariate method. This method handles heterogeneous and uncalibrated kerosene samples with optimal spectral profiles, which goes beyond quality assurance to target tax avoidance and adulteration at local filling stations and reduce refining process and crude oil source uncertainty by establishing effective kerosene compositional classification algorithms. These categorization algorithms ensure kerosene integrity, reliability, and regulatory compliance for industrial and home consumers.

Kerosene has been classified in only a few studies in the literature. To check for adulteration in kerosene samples, Pontes, et al. (2011) used PLS-DA and the successive projections method (SPA-LDA) in conjunction with near-infrared NIR data with an optical path of 10 mm. When it came to classifying a subset of the input data, PLS-DA produced the greatest results, with a 100% accuracy rate in the external set.

In regard to taxation and identification of kerosene adulteration at local fueling stations, the current study tested the multivariate methods (PLS-DA with HCA and PCA) for the direct classification of kerosene, which is a novel strategy never tried previously for kerosene, based on their conformance to the refining source, where different crude oil and refinery processes were used, from its optimized spectra profile. These classification models are necessary to overcome the difficulties of anomalous conditions generated from different kerosene samples; those were not previously included in calibration sets of existing prediction models.

II. Materials And Methods

A. Kerosene Sampling and Analysis

During a 6-month period, 60 kerosene samples were gathered from service stations in the eastern Iraqi city of Kalar to consider time-dependent variation in crude oil's chemical structure. The crude oil for these samples comes from Kirkuk and Bazian (Classes K and R, respectively), two of the largest local refineries with different process specifications. Their crude oil comes from different sources. Additional 10 kerosene samples were tested, obtained from a different source, outside the study area, for model verification (external validation). Polyethylene containers containing specimens for FTIR analysis were maintained at a temperature below 8°C in accordance with the usual procedure of the standard test method (ASTM D4057 - 19, 2019).

In this study, a spectrophotometer (Model: IRAffinity-1S: SHIMADZU) was used. The spectral resolution power was 4.0 cm^{-1}. For each sample, three absorbance spectra in the 4000–400 cm^{-1} wavenumber region were taken and averaged. Three spectra were multivariate calibrated to (ASTM E1655-17, 2018). Testing on the used spectrophotometer is being done using a routine quality assessment system developed according to standard procedure (ISO 4259-3, 2020). These measurement methods were evaluated for repeatability and reproducibility (accuracy) in accordance with (ISO 5725-2, 2019).

B. Multivariate Statistical Analysis and Model Development

The obtained FT-Mid IR spectral data was first pretreated for curve smoothing and baseline correction by applying the Savitzky-Golay first derivative method of a 21-point window and a second-order polynomial using OriginLab software (free trial version), resulting in 448 variables. The data were also preprocessed by changing the values to mean-centered – variance set to be 1. The results of each class were arranged into a matrix of samples as rows and spectra as columns. Kerosene samples were then subjected to statistical analysis using agglomerative hierarchal cluster analysis (AHC), which uses Euclidean distances and Ward linkage measures, to show that the samples were indeed grouped together. PCA analysis has been applied to each class to segregate the dataset into main groups (Issa, 2024). To further purify component scores, a varimax rotation of the PCs with significant eigenvalues was conducted. This allowed us to maximize the distribution of the components by minimizing the number of small coefficients whereas maintaining a high level of detail in the original data, as illustrated in Fig. 1a.

Next, samples have been divided into a calibration set after removing outliers, and a validation set using the Kennard-Stone algorithm (Kennard and Stone, 1969). The PLS-DA multivariate calibration method was used to establish the classification model with a threshold between 0 and 1. To verify an appropriate number of latent variables (LV) for the PLS-DA classification model, the cross-validated (leave one out) predictive relevance (Q^2) of the model that each LV manages to accomplish has been taken into account. A higher Q^2 value means a higher predictive ability for the model, as reported by Roy and Roy (2008).

$$Q^2 = 1 - \frac{\sum\left(Y_{meas} - Y_{pred.}\right)^2}{\sum\left(Y_{meas} - \bar{Y}\right)^2} \quad (1)$$

Outliers were detected using high leverage values and Q residuals at a 95% level of confidence. Sensitivity (Sens), specificity (Spec), precision (Pre), and accuracy (Acc), Equations 2–5, evaluations for PLS-DA model quality have been employed as reported by Mohammadi, et al. (2020) and presented in the following equations (2–5), where TP,

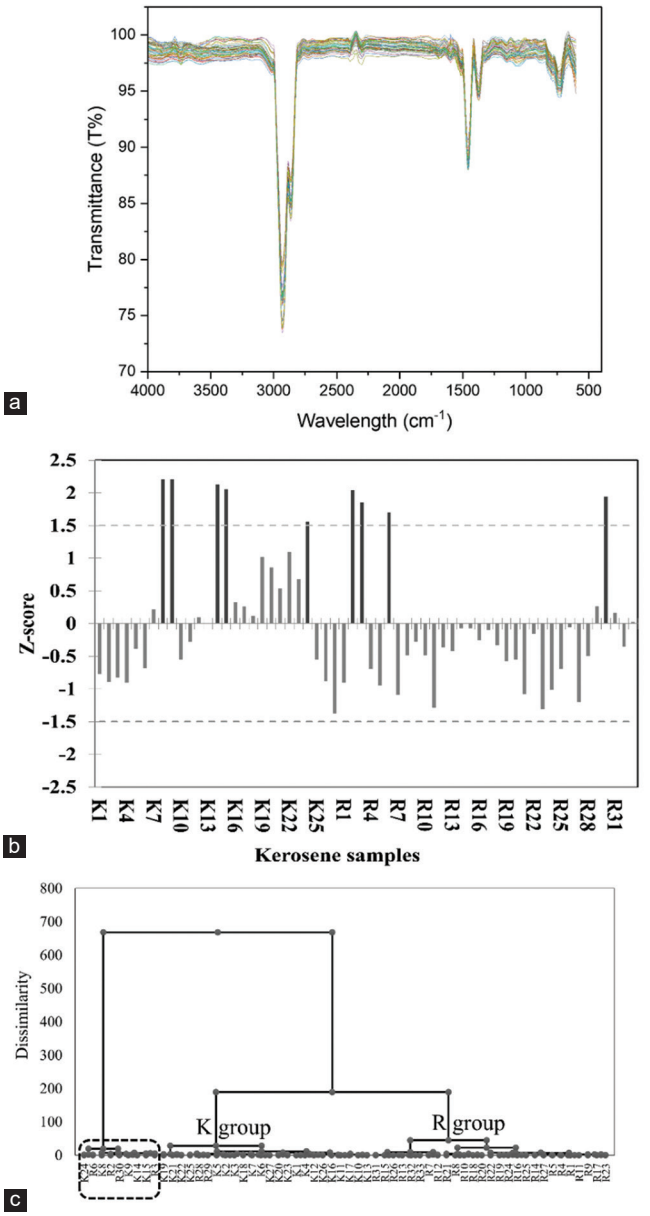

Fig. 1. (a) Mid-infrared spectrum of 60 kerosene samples in the range of 600–4000 cm^{-1}, (b) Results of Grubb's test for outlier detection at a significance level of 95%, and (c) Dendrogram of agglomerative hierarchal cluster for the 60 kerosene samples.

FN, TN, and FP represent the statistical parameters of true positive, false negative, true negative, and false positive, respectively.

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

$$\text{Pre} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

$$\text{Acc} = \frac{\text{TN} + \text{TP}}{(\text{TN} + \text{TP} + \text{FN} + \text{FP})} \quad (5)$$

III. Results and Discussion

A. Characteristics of Kerosene Samples

Fig. 1a presents the FT-Mid IR spectra analysis of kerosene samples after applying curve smoothing and baseline correction. It is apparent from Fig. 1a that kerosene samples are composed of several main hydrocarbon compounds. Several typical band vibrations within the FT-Mid IR range (4000–600 cm^{-1}) are presented. According to the existing vibration types in Fig. 1a, different stretching and bending of various functional groups for diagnostic and fingerprint regions have occurred within the tested range. The stretching vibrations of aliphatic and aromatic C-H, C=C, and C-C groups are presented. Bending vibration related to C-H also existed in tested samples around 730 cm^{-1} and 1035 cm^{-1}. In general, samples contain aliphatic compounds, with a considerable proportion of long linear types of absorption range around 730 cm^{-1}, and aromatic compounds with an absorption range of 1400–1600 cm^{-1} (Coates, 2000). Results of Grubb's test for outlier detection at a significance level of 95% are shown in Fig. 1b.

The AHC method was applied to determine if the unidentified outliers in Grubb's test should be rejected or kept in the dataset. According to the AHC method, results displayed in Fig. 1c, the questionable outlier samples existing in the dataset are clearly identified. These samples have demonstrated a high degree of dissimilarity when compared to the rest of the samples. These outliers, which are shared by both the R and K sample sets, were excluded in the next classification calculations. For the remaining dataset, the spectra of the samples that were measured within the FT-Mid IR range are correctly classified into two main clusters, which reflect the origin of the samples at the moment of sampling.

PCA was used to look at and sort samples of kerosene into different groups based on how they were refined and where they came from. As can be seen from the score plot of PCA results in Fig. 2a, Varimax rotation of the principal components (PCs) of significant eigenvalues was performed to explain more than 99.76% of the total variation in the IR spectra, which was represented by two main components, PC1 and PC2. PC1 (accounts 50.15%) and PC2 (accounts 49.61%) recognized two groups in the dataset corresponding to different sources and refining processes of kerosene samples.

A PLS-DA model has been established on the basis of two groups, R and K that were previously defined by PCA and AHC for FT-Mid IR spectral analysis of kerosene samples after excluding outliers, given that 36 samples were for the calibration set and 15 samples were for the prediction set. Of the 36 samples in the calibration set, 21 are of class R and 15 are of class K.

The 15 samples in the prediction set are comprised of 8 of R and 7 of K. The number of latent variables, for the supervised technique of PLS-DA, the term discriminant factor is also used for the spectral dataset was chosen on the basis of the maximum predictive ability explained for the dependent variable of the kerosene group classifier (Y) and FT-Mid IR explanatory variables Xi and using cross-validated Q^2.

http://dx.doi.org/10.14500/aro.11515

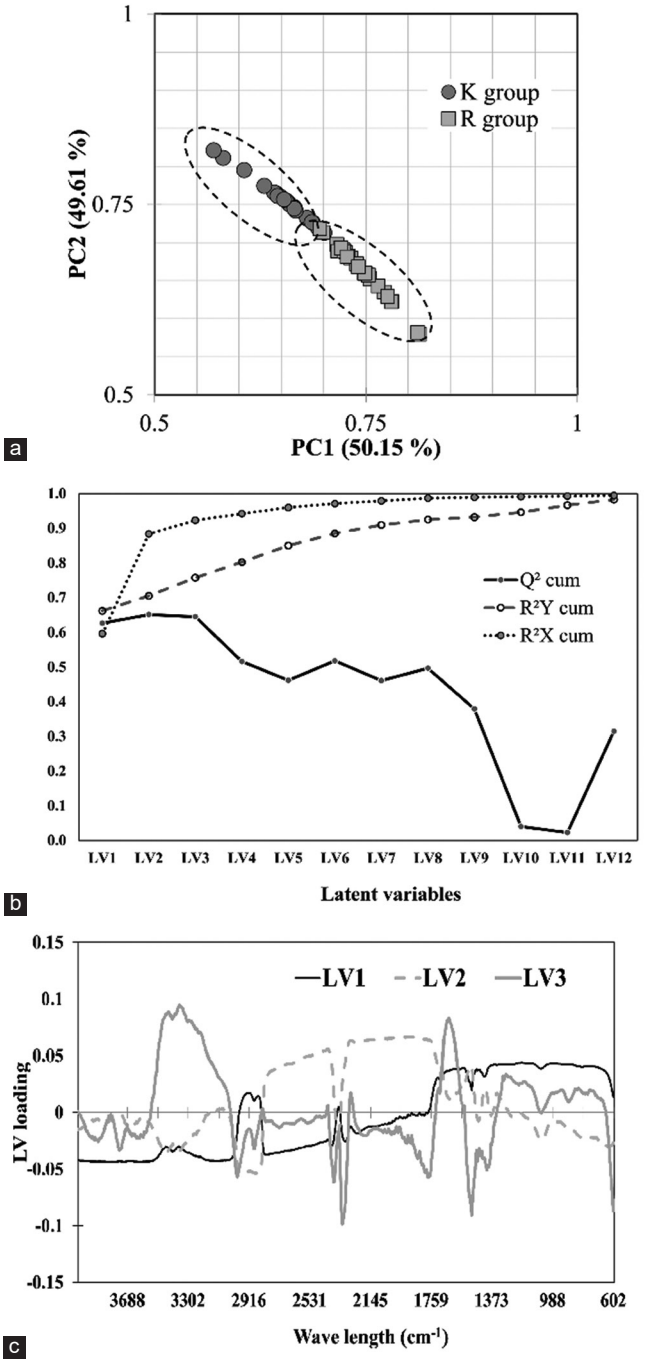

Fig. 2. (a) Principle component analysis score plots (PC1 and PC2) account 99.76% of the total variation in FT-Mid IR spectra for the kerosene dataset after Varimax rotation, (b) Plot of cross-validated Q2, R2Y, and R2X and number of LVs in partial least squares-discriminant analysis, (c) line loading plot for the three LVs for FT-Mid IR explanatory variables.

Fig. 2b shows that the cumulative Q^2 has been maximized when considering three LVs (Q^2 equals 65.13%) and that after this point, Q^2 values give lower predictive relevance for the model, despite the fact that a larger number of LVs succeeds in explaining a larger total variance of Y and Xi, as shown by the cumulative values of the regression coefficient (R^2). The total variance explained by three LVs for Y was 75.75% and for Xi was 92.29%.

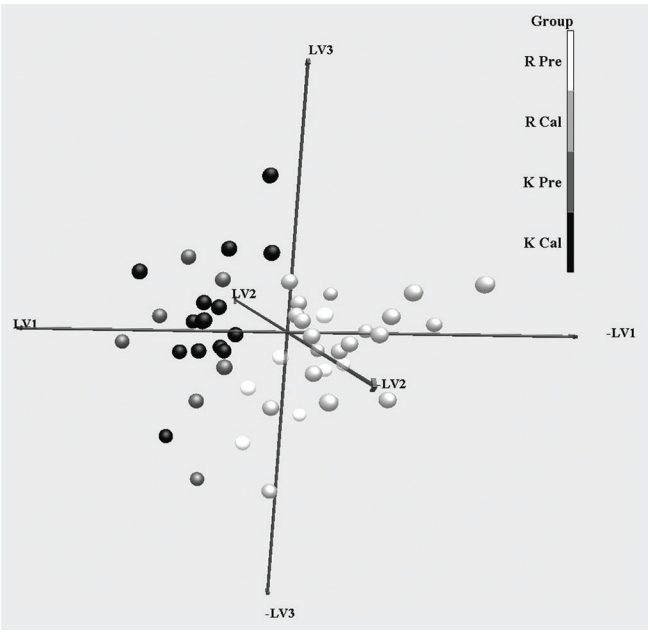

Fig. 3. Three-dimensional score plot of the LV1, LV2, and LV3 for calibration and prediction sets in partial least squares-discriminant analysis model.

The loading line for the LVs is depicted in Fig. 2c. It can be seen that the spectral transmission bands at wavenumbers between 600 and 1700 cm^{-1}, involving C-H and C-C bending and CH_3 stretch, are the sections that had the most impact on LV1. The absorption peaks at 2990–2400 and 2300–1650 cm^{-1}, which may have been caused by CH3 and C=C stretches, had a greater impact on LV2. In addition, the primary absorption peaks that characterized the LV3 are located between 3550 and 3028 cm^{-1}, 2450 and 2300 cm^{-1}, and 1700 and 1400 cm^{-1}, respectively, where these frequencies correlate to stretches in the CH_2 and CH_3 bonds, as well as aromatic C=C bonds.

The FT-Mid IR spectra shown in Fig. 2c, that the fingerprint region is the most significant for LV1. Using the PLS-DA model for three latent variables and leaving one out cross-validation technique, the data set was classified into two main groups. In Fig. 3, adequate separation performance for the samples is depicted by the 3D score plot of LV1, LV2, and LV3 (accounting for 59.63%, 28.74%, and 3.92% of the variances in the FT-Mid IR spectra, respectively). The figure shows that the PLS-DA model can correctly divide the kerosene samples into two Groups, K and R, based on how they were refined and where they came from, for the calibration and prediction sets.

Table I displays the results of a classification analysis of the PLS-DA model. The results demonstrate that 100% of the samples in the calibration set were correctly identified, whereas 86.67% of the samples in the prediction set were correctly classified. To further evaluate the efficacy of the PLS-DA model, the Sens, Spec, Pre, and Acc were computed for both the calibration and prediction sets, as shown in Table I. Sens and Spec are statistical parameters for measuring the dependability of any classification model, whereas Acc and Pre can be estimated to help with the realization of the model's specifications. The suggested

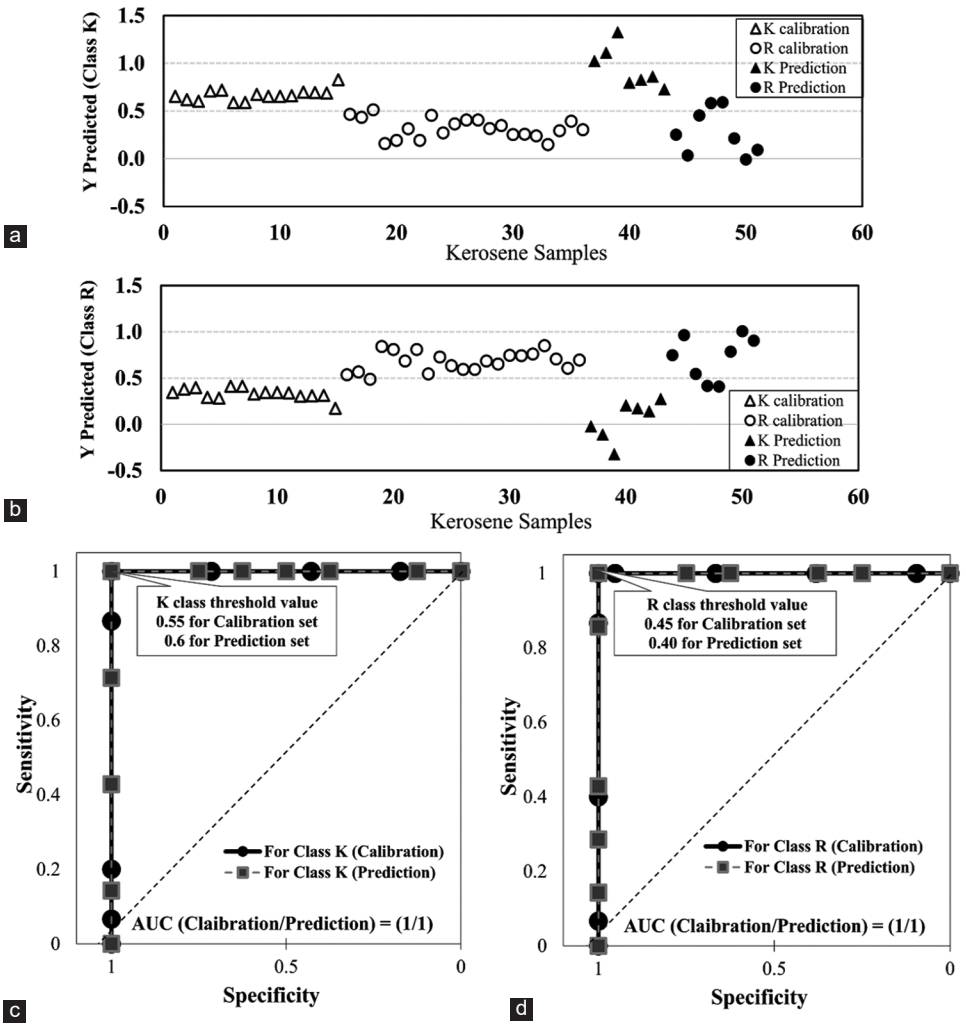

Fig. 4. (a) Class values estimation plots of calibration and prediction sets using partial least squares-discriminant analysis (PLS-DA) models for class K, (b) for class R, (c) ROC curve using PLS-DA models for classification of the kerosene samples for class K; and (d) for class R.

TABLE I
The Confusion Matrix And Classification Characterization of Kerosene Groups Classification using the PLS-DA Model for Calibration and Prediction Sets

Calibration set				
From/to	K	R	Total	% correct
K	15	0	15	100.00%
R	0	21	21	100.00%
Total	15	21	36	100.00%
Prediction set				
From/to	K	R	Total	% correct
K	7	0	7	100.00%
R	2	6	8	75.00%
Total	9	6	15	86.67%
Calibration set				
Class	Sens.	Spec.	Pre.	Acc.
K	1.00	1.00	1.00	1.00
R	1.00	1.00	1.00	1.00
Prediction set				
Class	Sens.	Spec.	Pre.	Acc.
K	1.00	0.75	0.78	0.78
R	0.75	1.00	1.00	0.87

PLS-DA: Partial least squares discriminant analysis

model's sensitivity measures the fraction of true positives and its specificity measures the fraction of false negatives for correctly classified data (Khanmohammadi, et al., 2013). Table I also shows that the model's quality evaluation demonstrates its dependability in classifying kerosene samples into their respective categories, with a Sens of 1.00 for class K and a Spec of 1.00 for class R in both the calibration and prediction sets. Model performance is slightly lower for the prediction set compared to the calibration set class for some defining parameters, which may be due to the relatively limited number of samples employed.

Fig. 4a and b display an analysis of the discrimination of kerosene samples using the PLS-DA model. The goal of this analysis is to determine the differences in refining procedures and crude oil origin that exist between the classes K and R that have been allocated to the FT-Mid IR spectral data set. The sample classification procedure appears to have been carried out satisfactorily, as evidenced by the accurate segregation of the class values for both the calibration and prediction sets. Fig. 4 demonstrates quite clearly how the two sets of kerosene samples, denoted as K and R, can be distinguished from one another.

The receiver operating characteristic (ROC) curve, which is utilized for the examination of classification abilities, has been applied to illustrate the performance, in terms of Sens and Spec, of an existing model as a function of varying the discrimination threshold (Hanley, 1998).

The ROC curves for the groups K and R that were examined are shown in Fig. 4c and d. It is clear that a threshold value of 0.55 or 0.6 for the K class will result in a sensitivity and specificity value of 100 percent. The same result has also been achieved for the R class, with threshold values of 0.45 and 0.4 for the calibration and prediction sets, respectively. This indicates that the two groups for K and R classes have been entirely separated from one another. Because the area under the curve (AUC) is always equal to 1.00, the p value is always less than 0.05, which indicates that the PLS-DA model diagnostic is significant. Since the PLS-DA model screening is significant, the null hypothesis H_o is rejected in this scenario. According to H_o, the area under the curve (AUC) is equal to 0.5, which indicates that the separation performance of the model is completely a matter of chance (Mallick, et al., 2022).

Before comparing the results of the PLS-DA model derived in this work to those of related previous studies, it is important to state that the results obtained by the PLS-DA model and those raised by PCA and AHC analysis methods are highly consistent, confirming the difference between the K and R classes of kerosene samples. This finding suggests that the differences in hydrocarbon composition between these two groups are the result of differences in refining processing and crude oil origin.

The external validation for the PLS-DA model was made up of validation samples of the classes R and K and 10 foreign samples (F) collected from a different source of kerosene samples, to check the model capability to discriminate the F samples as not belonging to any of the classes. Fig. 5 shows the observation chart, showing the distances of the F samples from the validation set of K and R classes. It can easily be seen that classes R and K are correctly classified and F samples are outside the boundaries of the studied classes K and R.

The PLS-DA model's verification was independently validated by an external source by distinguishing the 10 foreign samples (F), collected from a different source of kerosene, as not belonging to any of the classes was tested. Observation chart depicting F samples distances from the K and R classes in the validation set within the 5% confidence

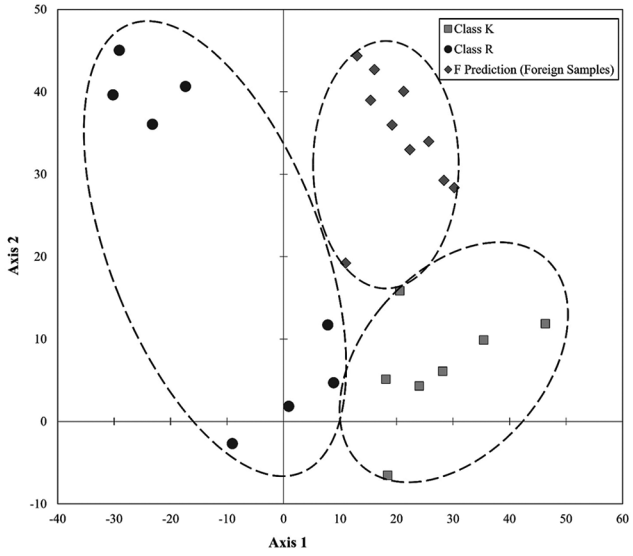

Fig. 5. Three classes observation chart with confidence ellipses for the partial least squares-discriminant analysis model external verification over the first two LVs (with a significance level of 5%).

TABLE II
SENS AND SPEC OF PLS-DA MODEL FOR KEROSENE SCREENING COMPARED TO PREVIOUS STUDIES.

Investigation approach	Ave. Sens[1]	Ave. Spec*	No. of LVs	No. of classes (NC)	NS [9]/NC
PLS-DA, IR (calib. set in this work)	1.00	1.00	3	2	18
PLS-DA, IR (pred. set in this work)	0.88	0.88			7.5
PLS-DA, IR for crude oil (calib. set)[2]	1.00	1.00	2	3	23.33
PLS-DA, IR for crude oil (pred. set)[2]	1.00	1.00			10
PLS-DA, GC-MS for gasoline (calib. set)[3]	0.97	0.99	3	4	16
PLS-DA, GC-MS for gasoline (pred. set)[3]	0.97	0.93			4
PLS-DA, IR for gasoline (calib. set)[3]	1.00	0.99	3	4	16
PLS-DA, IR for gasoline (pred. set)[3]	1.00	1.00			4
QDA, IR for gasoline (calib. set)[4]	0.90	0.90	3	2	20
QDA, IR for gasoline (pred. set)[4]	0.85	0.88			22.5
PLS2-DA, IR for biodiesel (calib. set)[5]	1.00	1.00	3	4	15
PLS2-DA, IR for biodiesel (pred. set)[5]	1.00	1.00			7
PLS-DA, FE for gasoline (calib. set)[6]	0.87	0.89	3	3	16.66
PLS-DA, FE for gasoline (pred. set)[6]	1.00	1.00			8.33
SIMCA, Phys Prop for kerosene (calib. set)[7]	0.79	0.29	2	2	16
SVM Phys Prop for kerosene (calib. set)[7]	1.00	1.00	-	2	20
SVM, Phys Prop for kerosene (pred. set)[7]	1.00	1.00			12.5
SIMCA, IR for kerosene (calib. set)[8]	1.00	1.00	2	2	20
SIMCA, IR for kerosene (pred. set)[8]	1.00	1.00			15

[1]Average of Sens and Spec values for classes were taken into account; [2]adopted from (Mohammadi, et al., 2020); [3]adopted from (Barra, et al., 2020); [4]adopted from (Khanmohammadi, et al., 2013); [5]adopted from (Mazivila, et al., 2015); [6]adopted from (De Paulo, et al., 2014); [7]adopted from (Comesaña-García, et al., 2013); [8]adopted from (Dago Morales, et al., 2008); 9NS is a number of samples. PLS-DA: Partial least squares discriminant analysis, SIMCA: Soft independent modeling of class analogy, SVM: Support vector machine

limits, Fig. 5 shows that over the latent variables LV1 and LV2. It is clear that classes R and K have been appropriately identified and that the F samples lie outside of the boundaries of the classes K and R that have been examined.

Table II displays a comparative analysis of the current study with different methods in the literature used to classify various fuels such as crude oil, gasoline, biodiesel, and kerosene. The models were established using techniques such as PLS-DA, quadratic discriminant analysis (QDA), support vector machine (SVM), and SIMCA, along with analytical methods such as infrared spectroscopy (IR), gas chromatography-mass spectrometry (GC-MS), and Physical properties (Phys Prop). Performance indicators such as Average Sensitivity (Ave. Sens) and Average Specificity (Ave. Spec) are provided for both calibration and prediction sets to demonstrate the models' capacity to generalize to new data. Some methods consistently perform well on both calibration and prediction sets, such as PLS-DA with IR for gasoline, crude oil, and biodiesel or SIMCA with IR and Phys Prop for kerosene. However, other methods show variability or lower performance on prediction sets compared to calibration sets. The outcomes obtained from this study, the developed PLS-DA model, for kerosene process detection indicate a potential model robustness and generalization. Despite differences, the comparison with previous models, the model here for kerosene has a promising accuracy in categorizing the samples, offering vital insights into their usefulness in fuel classification tasks.

The classification assessment presented in Table II reveals that despite the fact that chemometrics and spectral analysis have been utilized in numerous attempts to differentiate between crude oil and petroleum products, relatively little work has been done in this area with regard to the categorization of kerosene as it can be drawn from the researches in the literature. In the investigations conducted by (Comesaña-García, et al., 2013, Dago Morales, et al., 2008), the distinction between kerosene sample was achieved through the utilization of several modeling approaches. These approaches included SIMCA coupled with physical parameters and SVM coupled with IR spectra, there results were close to those obtained here, especially for the calibration data set.

This work utilizes a unique modeling strategy that combines PLS-DA with FT-Mid IR spectrum analysis to differentiate between two groups of kerosene based on their origins and refining processes. Throughout the investigation, it was shown that PLS-DA models can effectively do the task, leading to the highly satisfactory findings described earlier. The classification performance was promising despite the small sample size employed for the prediction set. The study used a limited number of samples, but the results demonstrate that PLS-DA is a valuable, statistically significant, straightforward, and cost-effective method for distinguishing kerosene, similar to its application for other petroleum products.

IV. Conclusions

- Utilizing the supervised pattern detection method PLS-DA and multivariate analyses of AHC and PCA on the FT-Mid IR dataset can aid in sorting kerosene from different refineries and sources efficiently for quality and taxation purposes.
- Kerosene samples from various routes can be effectively classified using PLS-DA.
- AHC has been applied professionally to classify kerosene samples and detect outliers.
- Outliers identified by Grubb's test were discarded.
- PCA with the varimax rotation method easily distinguished the sample distribution into two main classes.
- Three LVs from cross-validated PLS-DA models were utilized in the calibration set, resulting in successful discrimination.
- Compared to SIMCA and SVM models, PLS-DA demonstrated significant discrimination capability.
- The supervised PLS-DA discrimination model significantly improved the classification of kerosene samples into clear groups, achieving 100% accuracy in the calibration set and 86.7% accuracy in the prediction set.
- The study highlights the effectiveness of using supervised PLS-DA for sorting kerosene samples based on their origin and processing methods, facilitating quality control and fraud detection.
- Further research is needed to explore the combination effect of PLS-DA with SIMCA on classification accuracy.
- Other methodologies such as artificial neural networks and classification and regression trees could be promising for kerosene categorization and warrant investigation.

V. Acknowledgment

The authors express their gratitude to Lox-Agency Laboratories, Kalar-Garmian Region, for providing the required laboratories and facilities to complete the experimental portion of this study.

References

Abdullah, Z.F., and Daij, M.A., 2021. Analytical study of oil refining in the North refineries company for the period 2013-2019. *Journal of Al-Frahedis Arts*, 13(46), pp.131-159.

Ali, J.A., and Khodakarami, L., 2015. Optimal oil pipeline route in Kurdistan Region Taq Taq-Bazian refinery as case study. *International Journal of Engineering Trends and Technology*, 23(5), pp.257-262.

ASTM D4057 - 19, 2019. *Standard Practice for Manual Sampling of Petroleum and Petroleum Products*. ASTM International, West Conshohocken, PA.

ASTM E1655-17, 2018. *Standard Practices for Infrared Multivariate Quantitative Analysis*. ASTM International, West Conshohocken, PA.

Barra, I., Kharbach, M., Bousrabat, M., Cherrah, Y., Hanafi, M., Qannari, E.M., and Bouklouze, A., 2020. Discrimination of diesel fuels marketed in Morocco using FTIR, GC-MS analysis and chemometrics methods. *Talanta*, 209, p.120543.

Coates, J., 2000. Interpretation of infrared spectra, a practical approach. In: *Encyclopedia of Analytical Chemistry.* John Wiley and Sons Ltd., Chichester, UK.

Comesaña-García, Y., Cavado-Osorio, A., Linchenat-Dennes, E., and Dago-Morales, Á., 2013. Classification of kerosene using physicochemical data and multivariate techniques. *Revista CENIC Ciencias Químicas*, 44, pp.13-22.

Dadson, J., Pandam, S., and Asiedu, N., 2018. Modeling the characteristics and quantification of adulterants in gasoline using FTIR spectroscopy and chemometric calibrations. *Cogent Chemistry*, 4(1), p.1482637.

Dago Morales, Á., Cavado Osorio, A., Fernández Fernández, R., and Dennes, E.L., 2008. Development of a SIMCA model for classification of kerosene by infrared spectroscopy. *Química Nova*, 31, pp.1573-1576.

De Paulo, J.M., Barros, J.E.M., and Barbeira, P.J.S., 2014. Differentiation of gasoline samples using flame emission spectroscopy and partial least squares discriminate analysis. *Energy and Fuels*, 28(7), pp.4355-4361.

Hanley, J.A., 1998. Receiver operating characteristic (ROC) curves. In: Armitage, P., and Colton, Y., (eds.). *Encyclopedia of Biostatistics.* Wiley, Chichester.

ISO 4259-3, 2020. *Petroleum and Related Products - Precision of Measurement Methods and Results - Part 3: Monitoring and Verification of Published Precision data in Relation to Methods of Test*. ISO - International Organization for Standardization, Geneve.

ISO 5725-2, 2019. *Accuracy* (*Trueness and Precision*) *of Measurement Methods and Results - Part 2: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method*. ISO - International Organization for Standardization, Geneve.

Issa, H.M., 2024. Prediction of octane numbers for commercial gasoline using distillation curves: A comparative regression analysis between principal component and partial least squares methods. *Petroleum Science and Technology*, 42(10), pp.1233-1249.

Kaiser, M.J., 2017. A review of refinery complexity applications. *Petroleum Science*, 14(1), pp.167-194.

Kaltschmitt, T., and Deutschmann, O., 2012. Chapter 1 - Fuel processing for fuel cells. In*:* Sundmacher, K., (ed.). *Advances in Chemical Engineering.* Academic Press, Cambridge.

Karim, A.R., Khanaqa, P., and Shukur, D.A., 2017. Kurdistan crude oils as feedstock for production of aromatics. *Arabian Journal of Chemistry*, 10, pp.S2601-S2607.

Kennard, R.W., and Stone, L.A., 1969. Computer aided design of experiments. *Technometrics*, 11(1), pp.137-148.

Khanmohammadi, M., Garmarudi, A.B., Ghasemi, K., and De La Guardia, M., 2013. Quality based classification of gasoline samples by ATR-FTIR spectrometry using spectral feature selection with quadratic discriminant analysis. *Fuel*, 111, pp.96-102.

Lam, N.L., Smith, K.R., Gauthier, A., and Bates, M.N., 2012. Kerosene: A review of household uses and their hazards in low- and middle-income countries. *Journal of Toxicology and Environmental Health, Part B*, 15(6), pp.396-432.

Mallick, J., Talukdar, S., Alsubih, M., Ahmed, M., Islam, A.R.M.T., Shahfahad, and Thanh, N.V., 2022. Proposing receiver operating characteristic-based sensitivity analysis with introducing swarm optimized ensemble learning algorithms for groundwater potentiality modelling in Asir region, Saudi Arabia. *Geocarto International*, 37(15), pp.4361-4389.

Mazivila, S.J., Mitsutake, H., Santana, F.B.D., Gontijo, L.C., Santos, D.Q., and Borges Neto, W., 2015. Fast classification of different oils and routes used in biodiesel production using mid infrared spectroscopy and PLS2-DA. *Journal of the Brazilian Chemical Society*, 26(4), pp.642-648.

Mohammadi, M., Khanmohammadi Khorrami, M., Vatani, A., Ghasemzadeh, H., Vatanparast, H., Bahramian, A., and Fallah, A., 2020. Rapid determination and classification of crude oils by ATR-FTIR spectroscopy and chemometric methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 232, p.118157.

Naman, S., Jamil, L.A., Al-Gulami, F., Simo, S., and Ali, M.K., 2019. Evaluation of crude oils and natural gases of Kurdistan-Iraq by catalytic improvements to lighter oils using local clays. *WIT Transactions on Ecology and the Environment*, 222, pp.81-91.

Pontes, M.J.C., Pereira, C.F., Pimentel, M.F., Vasconcelos, F.V.C., and Silva, A.G.B., 2011. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification. *Talanta*, 85(4), pp.2159-2165.

Roy, P.P., and Roy, K., 2008. On some aspects of variable selection for partial least squares regression models. *QSAR and Combinatorial Science*, 27(3), pp.302-313.

Tanaka, G.T., De Oliveira Ferreira, F., Ferreira Da Silva, C.E., Flumignan, D.L., and De Oliveira, J.E., 2011. Chemometrics in fuel science: Demonstration of the feasibility of chemometrics analyses applied to physicochemical parameters to screen solvent tracers in Brazilian commercial gasoline. *Journal of Chemometrics*, 25(9), pp.487-495.