Identifikasi dan Prediksi Lokasi dan Pencitraan Bensin Segar Menggunakan Kromatogram Senyawa Bertarget Baru dengan Kemometri dan Pembelajaran Mesin

Identification and Prediction of Fresh Gasoline Locations and Branding Using Newly Targeted Compound Chromatograms with Chemometrics and Machine Learning

Aidil Fahmi Shadan,[1*] & Hafizan Juahir[2,3]

[1] Jabatan Kimia Malaysia, 46661 Petaling Jaya, Selangor, Malaysia.

[2] East Coast Environmental Research Institute (ESERI), Universiti Sultan Zainal Abidin, Gong Badak, 21300 Kuala Nerus, Terengganu, Malaysia.

[3]Faculty of Bioresources and Food Industry, Universiti Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia.

** Corresponding author, tel/fax +60129444666, email: aidilfahmi@kimia.gov.my*

Abstrak. *Deteksi dan penggunaan bensin di tempat kejadian criminal seperti pembakaran sangat diminati dalam penyelidikan forensik. Dalam karya ini, kromatografi gas-spektrometri massa (GC-MS) digunakan untuk menganalisis sampel bensin dan kemometrik yaitu analisis komponen utama (PCA), analisis diskriminan (DA), dan pembelajaran mesin klasifikasi dan pohon regresi (CART) diterapkan pada mengidentifikasi dan membedakan merek bensin dan lokasi asal. Studi ini mencakup tiga merek bensin populer yang dikumpulkan dari stasiun di delapan negara bagian Malaysia yang berbeda, termasuk satu kilang minyak. Hasil PCA dari 73,6% variasi komponen utama pertama dan kedua untuk kromatogram senyawa target baru (TCC) dan DA menggunakan metode analisis diskriminan dengan benar mengklasifikasikan 94,3% sampel pelatihan untuk lokasi asal dan 71,7% sampel pelatihan untuk merek. Model pembelajaran mesin two-C&R-trees (CART) baru juga dikembangkan dan diterapkan secara efektif pada 100 sampel bensin yang tidak dikenal, dengan rata-rata kesalahan absolut sebesar 1,1% (lokasi) dan 0,4% (merek). Hasil yang diperoleh menunjukkan potensi metodologi ini untuk membantu menyelesaikan investigasi kriminal.*

Kata kunci : ilmu forensik, pembakaran, teknik kemometri, Malaysia

Abstract. *The detection and use of gasoline at scenes of crimes such as arson is of high interest in forensic investigations. In this work, gas chromatography-mass spectrometry (GC-MS) was used to analyse the gasoline samples and chemometrics namely principal component analysis (PCA), discriminant analysis (DA), and classification and regression tree (CART) machine learning were applied to identify and discriminate the gasoline brands and locations of origin. This study includes three popular gasoline brands collected from stations in eight different Malaysian states, including one oil refinery. The PCA result of 73.6% variation of the first and second principal components for the new targeted compounds chromatogram (TCC) and DA using the discriminant-analysis method correctly classified 94.3% of training samples for location-of-origin and 71.7% of training samples for brand. A novel two-C&R-trees (CART) machine-learning model is also developed and effectively applied to 100 unidentified gasoline samples, with a mean absolute error of 1.1% (location) and 0.4% (brand). The obtained results demonstrate this methodology's potential to help resolve criminal investigations.*

Key words: forensic science, arson, chemometrics techniques, Malaysia

INTRODUCTION

The classification and detection of petroleum fuels is a vital part of the scientific investigation of arson, which has great significance for industrial, manufacturing, environmental, and forensic purposes. Arson can be described as a deliberate attempt to set a

property on fire or to eradicate evidence of a crime and is considered one of the easiest crimes to commit but one of the hardest to prosecute [1, 2]. Fuels include gasoline, kerosene oil, or diesel are commonly used as accelerants in illegal activities as they are readily available and cheap [2, 3]. International standards has been developed for the analysis and classification of Ignitable liquids also provide a general guide in forensic science for the definition and classification of the characteristics of petroleum products [4]. Quality control of fuels are accessed by technical specifications which can vary in different parts of the world (e.g., EN 228 in Europe, ASTM D48 14 in the USA, JIS K2202 in Japan and IS 2796 in India) [5].

Malaysia, which is geographically located near the equator, has hot and humid weather throughout the year. Thus, gasoline is very volatile and very difficult to detect in debris samples, especially in the summer. Samples sent to the laboratory will be analyzed using GC-MS and the chromatograms will be compared to those of gasoline. This process is time consuming and is subjected to individual interpretation. Besides the use of such processes, chemometric offer an alternative in the processing and classification of the results.

Several methods have been proposed for the detection and discrimination of fuel samples based on gas chromatography-mass spectrometry (GC-MS) [1–3, 6, 7], determination of additives by normal-phase high-performance liquid chromatography (NP-HPLC) combined with GC-MS [8], and various spectroscopic methods, such as nuclear magnetic resonance (NMR), near-infrared (NIR) spectroscopy, and attenuated total-reflection Fourier transform infrared (ATR-FTIR) spectroscopy [5, 9–13]. Although these analytical methods are widely used to determine the composition and classification of fuels, they have many drawbacks in that they are expensive to study. Many researchers have focused upon the determination of fuels using chemometric techniques such as principal-component analysis (PCA) [14], linear-discriminant analysis (LDA) [15], soft independent-class analog modeling (SIMCA) [16], and quadratic discriminant analysis (QDA) [17]. Unfortunately, there is no C&R trees (CART) modeling technique widely used in arson cases [18]. Ghazi et al. (2022), the sole Malaysian study on CART to date [19], reached the conclusion that the data requiring alignment prior to baseline correction or normalisation and the untargeted GC-MS data of neat gasoline should preferably be aligned first, followed by normalisation, and then by baseline correction. However, study adapting CART to real forensic cases are not highlighted as a validation of the developed model.

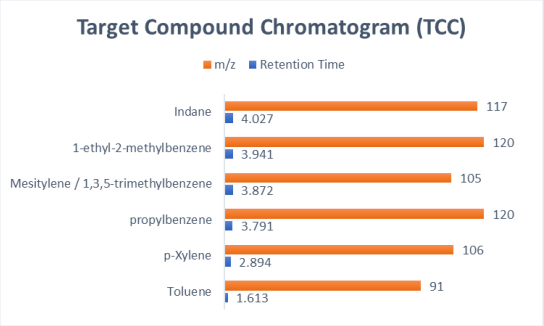

Figure 1. Target compound chromatogram (TCC) in this study.

Gas chromatography-mass spectrometry (GC-MS) was used to analyse the gasoline samples. The data was used to construct target compound chromatogram (TCC) which is toluene, p-xylene, propylbenzene, 1-ethyl-2-methylbenzene, mesitylene and indane (Fig. 1) prior to chemometrics. Principal component analysis was first employed to determine the compounds that has greater influence to the variance within a training data set. Factor analysis was used to determine the unique new targeted compounds used to differentiate gasoline brands and locations of origin. Discriminant analysis (DA) and C&R trees (CART) machine learning were applied to identify and discriminate the gasoline brands and locations of origin in this work.

MATERIALS AND METHODS

Materials

The present research involved the sampling of RON95 gasoline from the northern (Penang and Perak), western (Selangor and Kuala Lumpur), eastern (Pahang and Terengganu), and southern (Malacca and Johor) states of Malaysia (Table 1). From each state, three randomly selected commercial brands (P, S, and C) and one sample of gasoline from the Kerteh oil refinery were obtained as training data. Samples (100 mL) were directly obtained from pumps using separate amber-glass bottles and transported to the

laboratory at ambient temperature. Using the retention time based on premix hydrocarbon compounds, 100 samples of unknown brand and location and premix hydrocarbon were also analysed under the same conditions and used for cross-validation. At the laboratory, all samples were kept at 4 °C prior to analysis.

Table 1. Summary of the gasoline samples analyzed throughout this study.

State		Brand	Number of samples	Sample ID
West	Selangor	P	n = 3	B
		S	n = 3	
		C	n = 2	
	Kuala Lumpur	P	n = 2	W
		S	n = 2	
		C	n = 1	
North	Penang	P	n = 2	PP
		S	n = 2	
	Perak	P	n = 2	A
		S	n = 3	
South	Johor	P	n = 2	J
		S	n = 2	
	Malacca	P	n = 2	M
		S	n = 2	
		C	n = 1	
East	Terengganu	P	n = 6	T
	Pahang	P	n = 2	CA
		S	n = 2	
		C	n = 2	
	Kerteh (Refinery)	P	n = 4	TR
		S	n = 4	
		C	n = 2	
Unknown			n=100	UNK

Sample preparation

Gasoline samples were diluted 200 times with dichloromethane prior to analysis. The hydrocarbon test mix for fire-debris analysis (ASTM E1618) (AccuStandard) was also used to verify the retention time of compound of interest such as toluene and they are labelled as "RT-compound name" for the variables used in chemometrics and machine learning in all samples.

GC–MS analysis

All analyses were performed using an Agilent 7890A gas chromatograph coupled with an Agilent 5975C mass spectrometer (Agilent Technologies, Santa Clara, CA). The GC-MS was equipped with a HP5 fused-silica capillary column (30 m × 0.32 mm × 0.25 um, Agilent Technologies). Helium was used as a carrier gas at a nominal flow rate of 1.6 mLmin−1.; the inlet and transfer-line temperatures were held at 280 ◦C. The oven temperature was set to start at 40 ◦C (held for 2 minutes), then the temperature increased to 280 ◦C at a rate of 25 ◦C min−1 (held for three minutes), giving a total run time of 14.6 min. An electron-impact ionization source (70 eV) was utilized with a quadrupole mass analyzer operated in full-scan mode (m/z 40–550) at a sampling rate of 2.94 scans s−1. The sample-injection volume was 1 ml delivered by a headspace syringe with a split ratio of 20:1.

The compounds were identified by comparing the mass spectra with those spectra from National Institute of Standards and Technology mass-spectral search program Version 14, Gaithersburg, MD. The data were analyzed using the MSD CHEMSTATION Agilent Software. By employing the RTE Integrator Parameters, the obtained data have a minimum

peak area of 1,000, a centroidal peak position, a maximum of 60 peaks, and a 5-baseline reset point allocation. A compound was identified, and the peak area is used. Additionally, the selection of targeted-compound chromatograms applicable to the 53 samples was made using the guidelines provided by ASTM E1618-19 [20]. The use of statistical analysis in this study is very helpful for obtaining the uniqueness of the target-compound chromatograms.

Statistical analysis

The total area of each compound in the GC-MS chromatograms is exported to Microsoft Office Excel before being analyzed using XLSTAT factor analysis 2019.2.2 software.
Additionally, a total of 4 statistical modules were used, namely data preparation (transformation variables), data description (normality), data visualization (scatter plot), and data analysis (PCA and DA) for a total of 53 gasoline training data. 60 variables were utilised for PCA, and the implemented latent factor will consist of 6 variables. When matched to CRM, the 6 additional TCC variables are denoted as "RT-compound name" and referred to as supplementary variables. For discriminant analysis, 12 variables, 6 newly TCC and 6 RT from CRM, the same as those used for CART, were utilised. At an early stage in the determination of targeted-compound chromatograms in this study, factor analysis is used as a statistical module to obtain key latent factors.

The target compound chromatograms (TCC) is based on guidelines set out in ASTM 1618-19 [20], and a total of 100 samples of unknown gasoline brand and location of origin are used as prediction set to evaluate the efficacy of the CART model via the new TCC set.

RESULT AND DISCUSSION

Factor Analysis

The total area of each compound in the GC-MS chromatograms is exported to Microsoft Office Excel before being analyzed using XLSTAT factor analysis. A total of 60 variables for each of the 53 training samples provide a loading factor of 45.8% for the two main factors affecting the distribution of all variables. The main contributing variables can be determined as they have the largest cosine squared values in the sequence of factors, as well as in the patent factor table. Referring to the ASTM stated that 13 compounds would be useful for the classification of gasoline. Based factor pattern analysis, toluene, p-xylene, propylbenzene, and 1-ethyl-2-methylbenzene, mesitylene and indane are unique compounds which can differentiate the gasoline at factor analysis (Table 2).

Table 2. New target compounds throughout this study.

ASTM 1618-19		**THIS STUDY**	
Target Compound	**CAS NO**	**Target Compound obtained based on factor analysis**	**CAS NO**
Mesitylene/ 1,3,5-trimethylbenzene	000108-67-8	Toluene	000108-88-3
1,2,4-trimethylbenzene	95-36-3	p-Xylene	000106-42-3
1,2,3-trimethylbenzene	526-73-8	propylbenzene	000103-65-1
Indane	000496-11-7	Mesitylene / 1,3,5-trimethylbenzene	000108-67-8
1,2,4,5-tetramethylbenzene	000095-93-2	1-ethyl-2-methylbenzene	000611-14-3
1,2,3,5-tetramethylbenzene	000527-53-7	Indane	000496-11-7
5-Methylindane	874-35-1		
4-Methylindane	824-22-6		
Dodecane	112-40-3		
4,7- Dimethylindane	6682-71-9		
2- Methylnaphthalene	91-57-6		
1- Methylnaphthalene	90-12-0		
Ethylnaphthalenes (mixed)	1127-76-0		
1,3-Dimethylnaphthalene	575-41-7		
2,3-Dimethylnaphthalene	581-40-8		

After determining the 6 target compounds, all 53-training data were tested using statistical analysis. All variables first needed to be transformed using standardizing mathematical methods in the XLSTAT function (data preparation). The normality of the distribution of variables in each dataset is then tested. Using normality tests such as Shapiro-Wilk, Anderson-Darling, Lilliefors, and Jarque-Bera, all distributions of variables satisfied the Jarque-Bera test's alpha probability criteria of > 0.05, indicating that the data are normally distributed. The normality distribution of the training data was also visualized using the scatter-plot method; this test is very important for ensuring that these quantitative data are normally distributed.

Principal component analysis

Principal component analysis (PCA) which is an unsupervised model simplifies high-dimensional data while retaining its trends and patterns by transforming the data into fewer dimensions, which act as summaries of features. There was at least one correlation between the variables using Bartlett's sphericity test (significant level 0.05) and the Kaiser-Meyer-Olkin measure is larger than 0.5 and is thus acceptable so all variables can be used in the analysis. This value indicates that the total number of datasets in this study is sufficient.

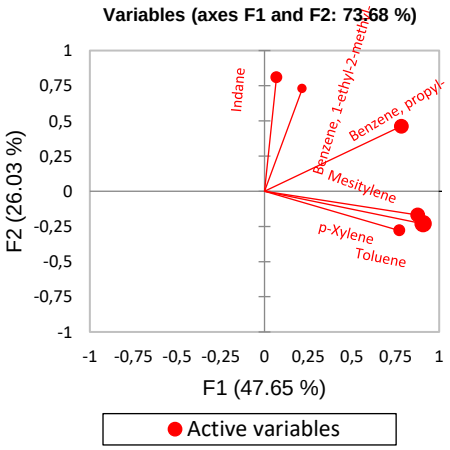

Figure 2. Biplot for active variables and observations. F1 is the first factor or principal component (PC1).

The PCA biplot showed that a total of 73.6% of variation in the samples was explained by of the first and second principal component. Indane, 1-ethyl-2-methylbenzene, and propylbenzene are the three main components with positive characteristics on both principles, while p-xylene, toluene, and mesitylene have a positive distribution on the first principle but a negative distribution on the second. The biplot more clearly shows the correlation between the active variables and the sample data (Fig. 2). The negative parts of the two principal components, which are based on the figure, are not represented by any of the active variables, but by the negative part of the first principle and the positive part of the second. This indicates that this PCA analysis requires supporting data to make it more accurate and reliable.

Since this study used the standard hydrocarbon premix solution to determine the retention-time of the compound of interest, the retention-time data were used to support PCA analysis by employing the supplementary data mode. The supplementary data consist of a total of sixty variables, six of which come from the premix hydrocarbons. Based on Fig. 2, the distribution of active and supplementary variables is better and covers each fraction of the two main components. The PCA biplot with extra supplementary data provides a clearer representation of the main components; therefore, with the availability of supplementary data, the results of PCA analysis can be clearly and easily understood.

Discriminant Analysis

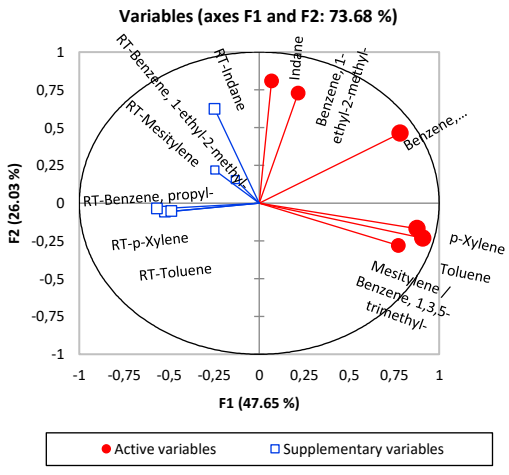

Figure 3. PCA plot for active and supplementary variables.

The gasoline training data used in this study include each sample's location and brand; these data can therefore be analyzed with the aim of discriminating them. A total of six newly

targeted compound chromatograms and six retention times (RT) of premix hydrocarbons are included in the training data that is utilised. Such analysis is well known and has been widely used in recent studies. The discrimination analysis method for finding the location in this study is to use the Pillai's test and the Hotelling–Lawley trace; this test will ensure that all data have unique mean vectors at significance level of 0.05. Location-discrimination analysis was able to discriminate 100% of the training data. Five main groups of sampling locations can be successfully discriminated by referring to Fig. 3.

Based on the results of the discriminant analysis of brands, Box test with the Chi-square asymptotic approximation method and Kullback's, respectively, showed the results that there was a difference in the covariance between each brand. Referring to the observation chart, the first factor (F1) represents 59.6% of the variance while F2 represents 40.3%of the variance, and it appears that three brands can be discriminated and that each has its own unique centroid (Fig. 4).

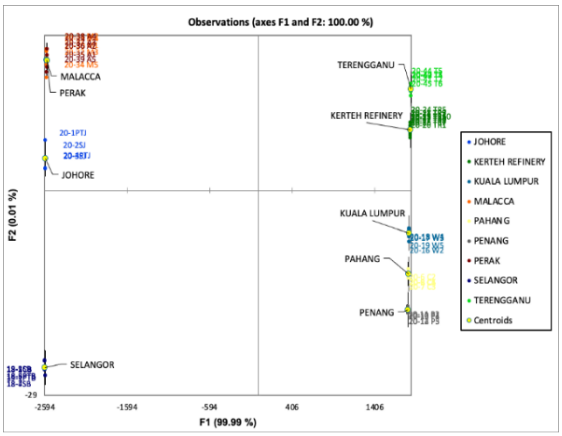

Figure 4. Discriminant analysis (DA) plot for location of the training data samples.

The confusion-matrix information for data training and cross-validation can also be explored using the discriminant-analysis approach. 94.3 % were correctly classified for location-of-origin sample training and 71.7 % correctly classified for brand using training samples. The cross-validation confusion matrix yielded results that were 94.3% correctly classified for location and 41.5% correctly classified for brand by using 100 samples as training data.; this means that the distribution of observational data for location is more scattered than that of brands, which is more converge and overlapping.

Classification and regression tree (CART) machine learning

Classification and regression tree (CART) machine learning belong to the family of supervised machine-learning algorithms [21]. In this analysis, the origin locations and brands of unknown gasoline samples can be determined based on the uniqueness of the newly targeted compound chromatograms using fifty-three training datasets. CART is based on a "divide-and-conquer" strategy whereby training data are subdivided into an increasingly pure and homogenous subset; the mathematical algorithm will therefore use six TCCs and six retention-time TCCs as possible predictors to obtain the key predictors to determine the locations or brands of unknown gasoline. Such primary predictors are at the top of the chart and are known as roots. The fractions of these main predictors are known as nodes or branches. Depending on the algorithm, nodes will end early and will most likely form new nodes below. If there is no node beneath the last node, it is referred to as a terminal node or leaf.

Referring to Table 3, a total of 100 unknown samples of gasoline were used to validate the CART model for prediction of location and brand. Nine samples were also to be used as validation samples. It is evident that the prediction standard deviation approaches the value of 0, indicating that the data are concentrated around the mean. The range of minimum and maximum values for training data and prediction data is comparable, with -4.9 to 4.7 for training data and -4.9 to 4.8 for prediction data, respectively. By using the CART method together with the GINI size mode, as many as 11 nodes and 8 rules will be generated with branches of three layers; this is in contrast to the CART model of the brand, whereby only 8 9 nodes and 6 rules with 8 layers of branches (Fig. 5) are produced. In decision tree construction, the principle of purity is determined by the proportion of the group's data elements that belong to the subset. The maximum percentage of purity in this study indicates that the rules in each node make accurate predictions. For example, 15.1% of training samples for node number 3 conform to the following algorithm rules: p-xylene > 1.10906, and brand P is the only comparable brand. This demonstrates that the classification at the end of this tree is one hundred percent pure.

Table 3. Training data set: a summary of unknown samples used to validate the CART machine-learning algorithm for predicting the brands and locations of origin.

Summary statistics (Training / Quantitative):

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Toluene	53	−2.171	2.911	0.000	1.000
p-Xylene	53	−1.985	1.610	0.000	1.000
Benzene, propyl-	53	−1.956	3.806	0.000	1.000
Mesitylene / Benzene, 1,3,5-trimethyl-	53	−1.096	2.648	0.000	1.000
Benzene, 1-ethyl-2-methyl-	53	−1.169	3.438	0.000	1.000
Indane	53	−0.721	1.846	0.000	1.000
RT-Toluene	53	−0.468	2.914	0.000	1.000
RT-p-Xylene	53	−0.408	4.757	0.000	1.000
RT-Benzene, propyl-	53	−0.361	2.812	0.000	1.000
RT-Mesitylene	53	−2.708	1.174	0.000	1.000
RT-Benzene, 1-ethyl-2-methyl-	53	−4.955	0.440	0.000	1.000
RT-Indane	53	−0.834	1.182	0.000	1.000

Summary statistics (Prediction set / Quantitative):

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Toluene	109	−1.063	3.244	−0.083	0.801
p-Xylene	109	−1.019	3.511	−0.079	0.774
Benzene, propyl-	109	−0.830	2.771	−0.041	0.694
Mesitylene / Benzene, 1,3,5-trimethyl-	109	−0.837	3.875	−0.048	0.834
Benzene, 1-ethyl-2-methyl-	109	−0.957	4.898	−0.003	0.992
Indane	109	−0.721	2.891	0.019	0.821
[#]RT-Toluene	109	−1.661	0.858	−0.049	0.970
RT-p-Xylene	109	−1.642	0.849	−0.052	0.971
RT-Benzene, propyl-	109	−1.628	0.723	−0.048	0.970
RT-Mesitylene	109	−2.708	1.141	−0.018	1.006
RT-Benzene, 1-ethyl-2-methyl-	109	−4.955	0.797	−0.051	1.077
RT-Indane	109	−1.800	1.175	0.039	1.015

[#] The peak identity of variable starting with the naming RT has been verified using the hydrocarbon test mix for fire-debris analysis

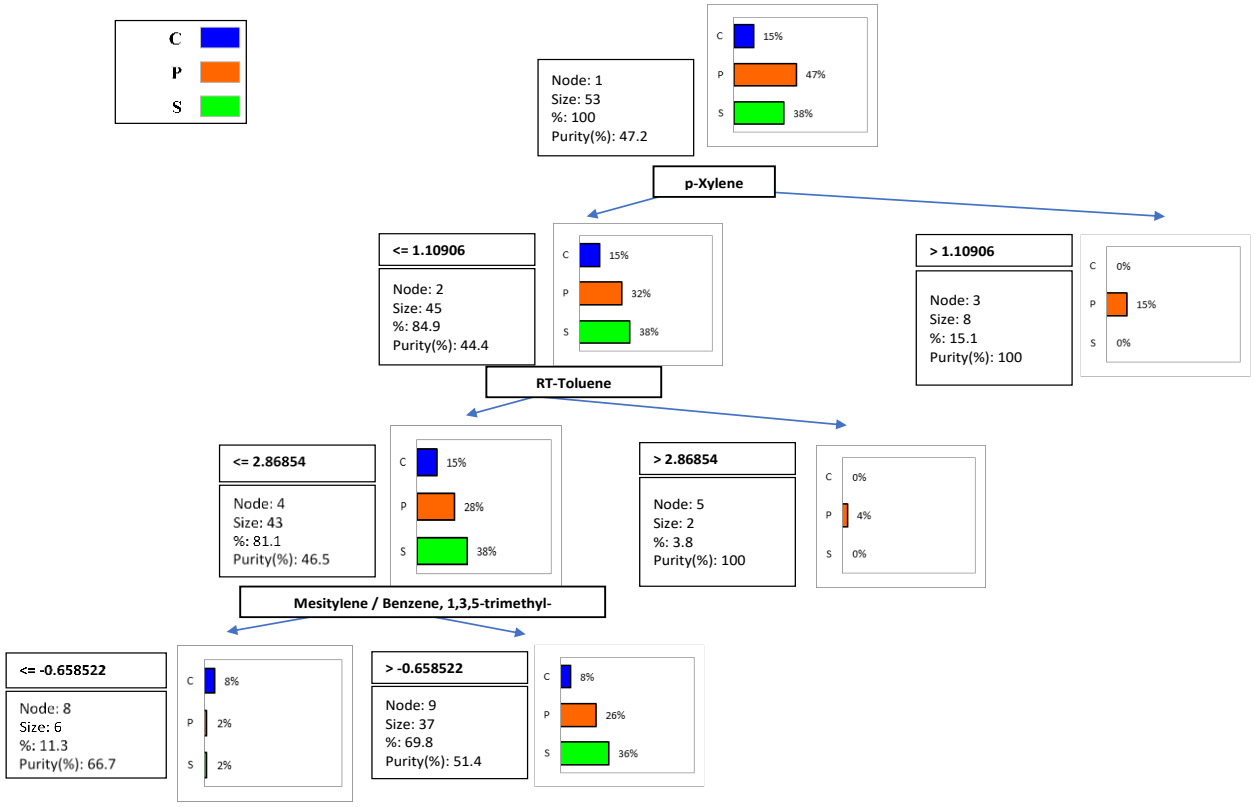

Figure 5. CART machine learning model for predicting unknown gasoline samples brand

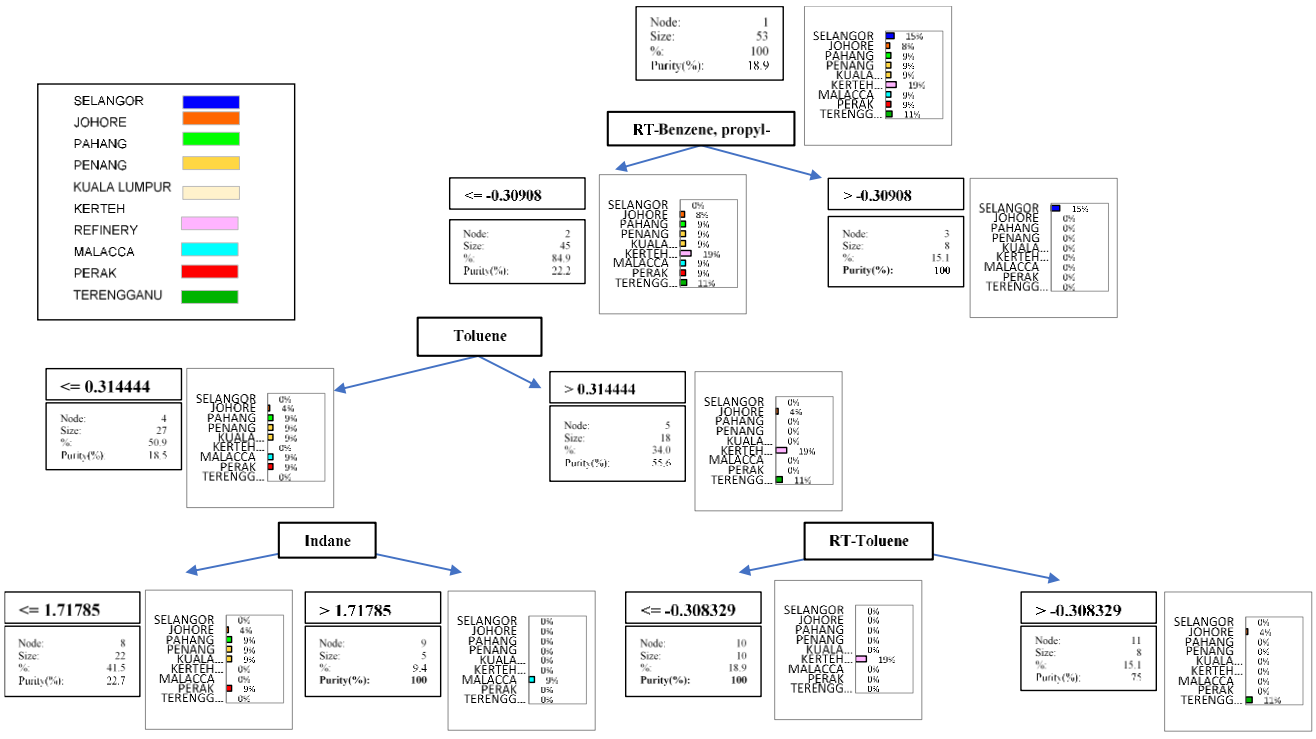

Figure 6. CART machine learning model for predicting unknown gasoline samples locations.

Similarly, as seen on the CART figure for location prediction, only nodes 9 and 10, which are located in Malacca and Kerteh, have a purity of 100 percent (Fig. 6). When interpreted based on main-location predictors, retention time dominates the key predictors for brand models; this shows that each model differs depending on the algorithm built by the key predictor.

Mean Absolute Error (MAE) scores are preferable the lower they are. This is due to the fact that MAE is a measure of the average error between predictions and intended targets, and we wish to minimise this value. The MAE for this study of the training data set was 1.1 (location) and 0.4 (brand), and the accuracy of the estimator was unaffected by modifications to the training data set. This study demonstrates that this method is ideally suited for use as Gonzalez et al. (2021) [22] did, utilising 243 unknown training data gasoline and achieving an MAE of 0.90 for their RON prediction model.

Although the two models had different rules, the locations of origin (Table 4) and the brands (Table 5) of all nine validation samples and 100 unknown gasoline samples could be predicted.

CONCLUSION

In conclusion, by changing the targeted compound chromatograms used as determining factors for the presence of gasoline such that they are appropriate to Malaysia, the brands and locations of origin of gasoline could be successfully determined using chemometric analysis and machine learning. For further studies, the addition of higher-volume data and certain other prediction factors—such as post-burning debris after exposure to weather, post-burning-debris sample duration, and exposure of debris to different environmental conditions—may be taken into account to achieve more comprehensive results for real case samples.

REFERENCES

1. Sinkov NA, Sandercock PML, Harynuk JJ. 2014, Chemometric classification of casework arson samples based on gasoline content. *Forensic Sci Int*.; 235:24-31.
2. Pert AD, Baron MG, Birkett JW. 2006, Review of analytical techniques for arson residues. *J Forensic Sci*.;51(5):1033–49.
3. Hupp AM, Marshall LJ, Campbell DI, et al. 2008, Chemometric analysis of diesel fuel for forensic and environmental applications. *Anal Chim Acta*;606(2):159–71.
4. Stauffer, Eric, Lentini, et al. 2003, ASTM standards for fire debris analysis: A review. *Forensic science international.;*132. 63-7.
5. Flumignan DL, Boralle N, de Oliveira JE. 2010, Screening Brazilian commercial gasoline quality by hydrogen nuclear magnetic resonance spectroscopic fingerprintings and pattern-recognition multivariate chemometric analysis. *Talanta*.;82(1):99–105.
6. González M, Ayuso J, Álvarez JA, et al. 2015, Application of an HS-MS for the detection of ignitable liquids from fire debris. *Talanta*.;142:150–6.
7. Peschier LJC, Grutters MMP, Hendrikse JN. 2018, Using alkylate components for classifying gasoline in fire debris samples. *J Forensic Sci*.;63(2):420–30.
8. Boczkaj G, Jaszczołt M, Przyjazny A, et al. 2013, Application of normal-phase high-performance liquid chromatography followed by gas chromatography for analytics of diesel fuel additives. *Anal Bioanal Chem.*;405(18):6095–103.
9. Balabin RM &Safieva RZ. 2008, Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. *Fuel*;87(7):1096–101.
10. Balabin RM, Safieva RZ, Lomakina EI. 2010, Gasoline classification using near infrared (NIR) spectroscopy data: com- parison of multivariate techniques. *Anal Chim Acta*;671(1-2):27–35.
11. Jais, Daud, Sharifah et al. 2020, Forensic Analysis of Accelerant on Different Fabrics Using Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy (ATR-FTIR) and Chemometrics Techniques. *Malaysian Journal of Medicine and Health Sciences.*
12. Pereira RCC, Skrobot VL, Castro EVR, et al. 2006, Determination of gasoline adulteration by principal components analysis-linear

discriminant analysis applied to FTIR spectra. *Energy Fuels*;20(3):1097–102.
13. Teixeira LSG, Oliveira FS, Dossantos HC, et al. 2008, Multivariate calibration in Fourier transform infrared spectrometry as a tool to detect adulterations in Brazilian gasoline. *Fuel*;87(3):346–352.
14. Malmquist LMV, Olsen RR, Hansen AB, et al. 2007, Assessment of oil weathering by gas chromatography-mass spectrometry, time warping and principal component analysis. *J Chromatogr A*.;1164(1-2):262–270.
15. González, M. J., Ferreiro-González, M., Barbero, et al. 2019, Novel method based on ion mobility spectrometry sum spectrum for the characterization of ignitable liquids in fire debris. *Talanta*. 199, 189–194.
16. Lee XQ, Sandercock PML, Harynuk JJ. 2016, The influence of temperature on the pyrolysis of household materials. *J Anal Appl Pyrol*.;118:75–85.
17. Figueiredo, Christophe B.Y. Cordella, Delphine Jouan-Rimbaud Bouveresse, et al. 2018, Evaluation of an untargeted chemometric approach for the source inference of ignitable liquids in forensic science, *Forensic Science International*.
18. Kumar R, Sharma V. 2018, Chemometrics in *Forensic Science, Trends in Analytical Chemistry*.;105:191-201.
19. Ghazi MGM, Lee LC, Samsudin AS et al. 2022, Evaluation of ensemble data preprocessing strategy on forensic gasoline classification using untargeted GC–MS data and classification and regression tree (CART) algorithm, *Microchemical Journal*;182.
20. ASTM E1618-19. 2019, Standard test method for ignitable liquid residues in extracts from fire debris samples by gas chromatography-mass spectrometry, ASTM International, West Conshohocken, PA, *www.astm.org. http://www.astm.org/cgi-bin/resolver.cgi?E1618-19*
21. Steinberg D. Chapter 10 CART: Classification and regression trees. 2009. *https://www.semanticscholar.org/paper/Chapter-10-CART-%3A-Classification-and-Regression-Steinberg/53dd76e162f69c48652a1146b32dd5c06792a801*
22. Gonzalez S, Kroyan Y, Sarjovaara T et al. 2021, Prediction of Gasoline Blend Ignition Characteristics Using Machine Learning Models, *Energy and Fuels*;35: 9332–9340.