Support vector machines for oil classification link with polyaromatic hydrocarbon contamination in the environment
Azimah Ismail, Hafizan Juahir, Saiful Bahri Mohamed, Mohd Ekhwan Toriman, Azlina Md. Kassim, Sharifuddin Md. Zain, Hadieh Monajemi, Wan Kamaruzaman Wan Ahmad, Munirah Abdul Zali, Ananthy Retnam, Mohd. Zaki Mohd. Taib, Mazlin Mokhtar and Siti Nor Fazillah Abdullah

ABSTRACT
The main focus of this study is exploring the spatial distribution of polyaromatics hydrocarbon links between oil spills in the environment via Support Vector Machines based on Kernel-Radial Basis Function (RBF) approach for high precision classification of oil spill type from its sample fingerprinting in Peninsular Malaysia. The results show the highest concentrations of Σ Alkylated PAHs and Σ EPA PAHs in ΣTAH concentration in diesel from the oil samples PP3_liquid and GP6_Jetty achieving 100% classification output, corresponding to coherent decision boundary and projective subspace estimation. The high dimensional nature of this approach has led to the existence of a perfect separability of the oil type classification from four clustered oil type components; i.e. diesel, bunker C, Mixture Oil (MO), lube oil and Waste Oil (WO) with the slack variables of ξ ≠ 0. Of the four clusters, only the SVs of two are correctly predicted, namely diesel and MO. The kernel-RBF approach provides efficient and reliable oil sample classification, enabling the oil classification to be optimally performed within a relatively short period of execution and a faster dataset classification where the slack variables ξ are non-zero.

Key words | classifier, fingerprints, oil classification, support vector machines

HIGHLIGHTS
- Support vector machines (SVM) approach is a reliable high precision classification technique.
- High dimensional nature lead to perfect oil types separability.
- This technique provides faster datasets classification.
- It leads towards a faster decision making to perform water cleaning process.
- This approach enables the oil classification to be optimally performed within a relatively short period of time.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/).
doi: 10.2166/wst.2021.038
INTRODUCTION

Millions of marine species that are inhabiting the oceans can be endangered if fuel oil is released into the sea or ocean (Gaganis & Pasadakis 2006). Even though no direct aggravation or deleterious health consequences are presented to the human population, oil spills can still cause damage to the coastal environment such as fish stocks or coral reefs. However, the wind and currents can push the oil strands towards the shorelines, which would result in contaminating the human habitat. This occurs particularly when the physical and chemical properties of the oil degrade during the severe environmental weathering process (Ramsey et al. 2014). In Malaysia, there are many fuel-oil spill-related issues that pollute Malaysian waters and land. These oil spills can be caused by refinery petroleum plants and waste oil amongst other small or large-scale oil spills (Ismail et al. 2016; Juahir et al. 2017). The small-scale oil spills refers to the small volume oil spills (e.g at the fishermens’ area), whereas the large-scale oil spills refer to large amounts of oil spill (e.g from the collision of ships or tankers). For example, the deadly collision of oil tankers in the shipping waterway that occurred on August 19, 2009 in the Straits of Malacca resulted in 58,000 tonnes of naphtha oil spilling into the ocean (Ahmad 2009). Moreover, the oil spills can also be from the cars, marine engines and deliberate oil discharge into the environment. The oil spills that have occurred in the Straits of Malacca, Johor Straits as well as South China Sea are caused by various reasons including accidental oil spills, intentional discharges from fishermens’ boats and maintenance and cleaning activities from the cargo vessels.

In order to implement an effective cleaning project, one needs to determine the oil type based on its polyaromatic hydrocarbon components. This requires a thorough oil spill sampling in the seawater surface along the coastal lines of Peninsular Malaysia since their presence in the environment tends to be rather recalcitrant (Ramsey et al. 2014). Performing oil spill sampling on a large amount of petroleum hydrocarbon oil spills in the oceans, seawaters, rivers, drains, water bodies and soil requires several strategic oil type classification methods that are very expensive and time consuming. The high complexity of the data requires an in-depth study on the restructuring of the data classification into the desired categories that is in line with the global scale in terms of competitiveness (Alamdar et al. 2016). Using innovative methods such as the gas chromatography- flame ionization detector (GC-FID) and gas chromatography-mass spectrometry (GC-MS) enables the oil spill fingerprinting to be used as source identification, oil spill characterization and an environmental forensics technique. Through utilizing analytical development laboratory tools for highly complex samples, one can determine the PAHs of major petroleum-specific targets that need to be chemically characterized, enabling the identification of the origin of the oil spill in a sample (Ismail et al. 2018). Furthermore, using the GC-FID and GC-MS methods in the oil spill fingerprinting study enables the comparable or
comparative analysis of the data with the unknown sources of the oil spills. Laboratory analysis alone for characterizing and identifying the complexity of petroleum hydrocarbon oil spills throughout the samplings and towards a result interpretation that might lead to measurement errors.

In this study, we employed the Support Vector Machines method, a fundamental approach which has offered an accurate classification technique and data prediction, and is a similar concept to artificial neural networks (Amjady et al. 2010). The building of substantial non-linear classification boundaries was performed by the kernel function of SVM, as an alternative measure in pattern recognition (Dufrenois & Noyer 2016). This learning machine enables the improvement of the output results of GC-FID and GC-MS to achieve the best oil type classification and higher profit with no unstable patterns (overfitting). Unlike ANNs, the SVMs complexity algorithm is an independent input space dimensionality that often outperforms ANNs in practice. Furthermore, it is less prone to overfitting when we are dealing with the large complexity dataset of oil spills.

The adoption of the SVM technique provides a strong support that has a benefit in decision making to gain the best results among electricity market players (Silva et al. 2017). There is however no method to apply SVMs in any oil spill fingerprinting area in Malaysia. In general, SVMs have brought the most reliable improvement in pattern recognition for smoothing large or complex datasets of clustering and regression in this study. This paper presents SVMs based on the Kernel – Radial Basis Function (RBF) algorithm, which was applied as a development approach for oil spill type classification using GC-FID and GC-MS. The objective of introducing SVM RBF in oil spill fingerprinting is to provide various statistical-based problem solving suggestions for obtaining oil spill type classification with a level of precision that can recognize the most significant petroleum-hydrocarbon compounds. Hence, the application of SVM RBF as the decision-making function is an innovative approach to obtain highly reliable output within a short execution time. The discovery of the maximal margin hyperplane enables a linear learning separation of the computational complexity in the non-separable datasets. Practically, the application of SVM is diverse in many fields, such as computer-based programming. The SVM RBF approach has high advantages of generative production data with no implicit assumption to ensure that the optimal classification boundary is achieved and the associated risks are minimized (Dufrenois & Noyer 2016). The function of SVM RBF in oil classification from oil spill fingerprinting is to improve the accuracy of boundary classification of complex mixtures of datasets and discriminate the most significant petroleum-oil type categories. In this study, we have explored a wide range of applications considering the RBF-SVM model as a decision function classifier to gain high precision of oil type classification performance from the oil spill fingerprints. In addition, this study provides an alternative approach to achieving significantly high oil classification accuracy of the potential sources that would lead to water pollution within Peninsular Malaysia.

**METHODOLOGY**

**Sample collection**

The oil samples were collected from the water surface in the form of oil films or sheens where the oil-contaminated water (liquid-liquid) as well as Teflon mesh fabric net (net-oil) (10 cm × 10 cm) were used. The samples were collected from the ports, jetty and fishermen’s areas, as illustrated in Figure 1 and shown in Table S1. The number of sampling points was based on site observation of the spill areas or the areas covered with oil, where for the bigger spill area, more sampling points were used. The samples were collected from February 2016 until September 2016. All oil samples were poured into 100–250 ml thick-walled wide neck borosilicate glass bottles with an inner neck diameter of 30 mm. The bottles were properly sealed and placed in a cool box. The temperature was maintained at +4 °C for delivery to the laboratory, where all the samples were kept at constant temperature of 4 °C until further analysis.

**Samples preparation procedure**

The oil sample was added with 60 ml Dichloromethane (DCM), which was subsequently shaken in the flask for complete oil extraction for 20 minutes. The procedures were performed in triplicate. Then, the oil sample was passed through the funnel containing anhydrous sodium sulfate (5.0 cm thick), and the round bottom flask was used to collect the dried extract at the end of the separating funnel. The 60 ml of DCM was further added into the separating funnel for further extraction of the oil from the water. The dried extract was finally concentrated into 1.0 ml in a rotary evaporator (30 °C) using the solvent-exchanged with hexane.
Column cleanup

A pre-cleaned glass column (30 cm × 10.5 mm I.D.) was used to perform the column cleanup analysis. It contains about 6.0 g pre-cleaned silica gel (100–200 mesh, Davisil grade 923) fused with 0.5 cm pre-cleaned sodium sulfate conditioned with 20 ml hexane. Initially, the extract was spiked with surrogates (4 mix PAHs compounds and α-terphenyl). 3.0 ml of hexane was repeatedly added onto the column for oil transfer into the column, followed by adding the 12.0 ml of hexane to elute aliphatic compounds. The aliphatic compound consists of n-alkanes and biomarkers (represents of F1) for further used for analysis of the saturates and biomarker. Secondly, to elute the aromatic compounds represents the F2, the 15 ml of 50% DCM in hexane (50:50 of dichloromethane: hexane) was used. A gentle stream of nitrogen (N₂) (1.0 ml/s) was used to concentrate both mixtures of F1 and F2 to reach the volume of less than 0.5 ml. The spike process procedures proceeded with F1 spiked with 1.0 ppm of C3017 β(H), 21β(H)-hopane and 20 ppm of 5α-androstane, and F2 spiked with 1 ppm of d14-terphenyl. The mixture of F3 was formed from the combination of the remaining half of F1 and F2 to determine of the total GC-detectable TPH, GC-resolved peak, and GC the unresolved complex mixture of hydrocarbon (UCM). Prior to analysis, all the three fractions underwent the concentration process to the final injection volume of 1.0 ml.

GC-FID and GC-MS analysis of the oil spill fingerprints

The laboratory analysis of GC-FID and GC-MS approaches are rather important in identifying the source, type and distribution patterns recognition of the oil spills. The analyses
of petroleum biomarkers of \( n \)-alkanes and PAHs in complex hydrocarbon mixture oil samples are performed based on the methods elaborated by (Wang 2003; Wang et al. 2005; Khelifa et al. 2007). Subsequent analyses for both \( n \)-alkanes distributions and total petroleum hydrocarbons (TPHs) of oil spilled clustering or separation were performed on a Perkin Elmer, Clarus 680 equipped with PE AutoSystem GC with built-in Autosampler and flame ionization detector (FID) on column dimensions of 30 m × 0.25 mm ID DB-5. The Agilent, 7890A GC System, which is equipped with a mass-selective detector and CTC PAL ALS Autosampler, was used to perform the analysis or identification of target PAH compounds (including five alkylated PAH homologous groups and other EPA priority PAHs) and high-molecular-weight biomarkers such as terpanes and steranes. A 30 m × 0.25 mm HP-5MS fused silica column was used for identifying the target PAH compounds. The common practice of laboratory analysis using the repeatability limits (three times) for substantial reliability of the results was carried out. The repeatability of the analysis has resulted in very similar concentrations, where the measurement was calculated based on the mean. All the clustering oil types from the collected samples of the GC-FID and GC-MS were proceeded with the support vector machine for the result confirmation. The approach tried applied in this study is illustrated in Figure 2.

**Support vector machines – kernel based radial basis function (RBF) approaches in oil spills classification**

**Classification support vector machines (CSVM) kernel – radial basis function (RBF)**

In this study, the CSVM Kernel-RBF approach was employed to validate the results obtained from the laboratory analyses by the gas chromatography – flame ionization detector (GC-FID) and gas chromatography mass spectrometer (GC-MS). The application of CSVM in oil spill fingerprints attempts to learn the oil spill datasets by finding the optimal margin hyperplane for perfect separation or classification of the large dataset from different physico-chemical characteristics (oil spill types, e.g. diesel). The CSVM analysis was performed using Statistica 13 Dell software 2015. The datasets were partitioned into three procedures: testing, training and predicting. In the common domain, the SVM classifier trains the datasets and builds

![Flowchart of oil spill identification process (Daling et al. 2002).](http://iwaponline.com/wst/article-pdf/83/5/1039/857870/wst083051039.pdf)
the hyper-plane through the Cutting Plane Algorithm (CPA) as decision boundary functions of separating the finite datasets with the margin maximized. The training process of oil spill datasets enables use of the hyperplanes found to estimate which oil type the datasets belong to. The discriminative power between groups during the training process provides perfect separation of the complex oil spill data sets into the respective groups. The validation approach was also accomplished using the SVM Kernel-Radial Basis Function (RBF). The execution of CSVM Kernel-RBF was based on Classification Type 1 and the Kernel-based approach, which is reliable for a non-linear boundary (Lavine et al. 2001; Dufrenois & Noyer 2016).

In this study, the forty-seven datasets were randomly trained and validated. In the Kernel-RBF approach, the hidden layer provides a set of functions, forming an arbitrary basis as the decision boundary for oil spill compounds. They are then converted as an input layer for the input vectors \((x_1, x_2, ..., x_{m_o})\), which are then subsequently expanded or spanned through the hidden layer of kernel inner product (support vectors \((x_i)\)), where the linear output layer maps the input vectors (training vectors) to optimal classification as the output \((y)\). Kernel inner product provides the optimal hyperplane construction or a similar function of the neuron network in the Artificial Neural Network (ANN) in the output feature space. The SVM architecture functioning as the decision functions is illustrated in Figure 3. During the training level, the proper weights and Support Vectors (independent variables) are properly computed that would generate non-zero, \(\alpha\) co-efficient which is then considered in the calculation (Haykin 1994; Singh et al. 2011).

(ii) Classification and Regression Models
(a) Data Pre-processing

The complete dataset of 47 primary data and 32 secondary data from the input data comprising 94 oil spill compounds are modeled in this study. Equation (1) was used to normalize the input dataset (Adib et al. 2013) to the dataset bounded region of \(y\in(+1,-1)\) to prevent truncation error as large amounts of numerical oil spill dataset:

\[
X_{ni} = \frac{2(X_i - X_{min})}{X_{max} - X_{min}} - 1
\]

where \(X_{ni}\) is a set of scaled input or output oil spill data, \(X_i\) is the actual input/output oil spill data, \(X_{min}\) is the minimum value of observed data set, and \(X_{max}\) is the maximum value of the observed dataset. Since the input dataset \(\langle X_{ni}\rangle_{n=1}^{N}\) was normalized using Equation (1), the output model of \(y_i\) was also normalized. The input dataset of oil spill was mapped into the normalized feature-space according to Equation (1) before running the data.

The normalized dataset of \(\langle X_{ni}\rangle_{n=1}^{N}\) was randomly divided into three subsets, namely training, optimization and testing (Adib et al. 2013). In this analysis, 80% of the data was used for the subsets of the training, 10% for optimization and another 10% for testing. The details of the oil spill datasets used for three subsets are elaborated in the next section.

(b) Data Training Process

Phase 1: The training dataset was given using \(\langle X_{ni}\rangle_{n=1}^{N}\) formula, where \(N\) represents the number of oil spill samples.
and $X \in \mathbb{R}^n$ is the input vector of $n$ input-features and \( \{y_n\}_{n=1}^N \) is the output feature with class labels or bound region $y_i \in \{-1, 1\}$ as the finite value. In binary classification, the function $y_i \in \{-1, 1\}$, however, $y_i \in \{1,2,3,4,5,6,7,8,9,0\}$ (Rodriguez-galiano et al. 2015). Prior to training the dataset, the optimum values of key parameters ($\gamma$, $\epsilon$, $\sigma^2$) were selected, where $\gamma$ is used for the fitting error minimization and the estimated function smoothness, $\epsilon$ is referred to as the precision threshold and $\sigma^2$ determines the efficiency of the SVM model. The slack variable, $\xi$ was also used to quantify the output features that consider the positive and negative classes (Adib et al. 2013) (Equation (2)):

\[
y(n) - f(Xn) \leq \xi n + \epsilon
\]

\[
f(Xn) - y(n) \leq \xi n + \epsilon
\]  
(2)

\[
\epsilon, \xi n, \xi n + \geq 0, n
\]

The $\epsilon$ in Equation (2) is the sensitivity of the optimal misclassification error, and \( \{\xi n\}_{n=1}^N \) are the slack variables to quantify the output features that consider the positive and negative classes. The SVM model works for the maximum-margin hyperplane where the separating hyperplane can be written as Equation (3) (García Nieto et al. 2013):

\[
y_i = w \cdot x_i + b
\]  
(3)

where $x_i$ ($x_1, x_2, x_3, \ldots x_n$) denotes the real-input of standard scalar dimensional vector, $w$ is the normal real-vector to a high-dimensional (hyperplane) feature space, and $b$ is known as a hyper-plane bias or the offset of the hyperplane from the origin line. For the case of linear separable, two hyperplanes can function as Equations (4) and (5) below in order to prevent any data point falling between the hyperplanes (García Nieto et al. 2013):

\[
Wx_k - b = 1
\]  
(4)

\[
Wx_k - b = -1
\]  
(5)

**Phase 2:** The SVM was then applied to find the functional form of classification and regression to deal with the non-linearity separation, which is subject to misclassification error function or constraint minimization. The training data in the SVM model includes the sequential error function optimization. Two SVM models were selected to solve the oil spill compound problems.

(i) Classification SVM (C-SVM classification): For this C-SVM, the error function minimization was resolved where the equation is later trivially defined as (García Nieto et al. 2013):

\[
\frac{1}{2} W^T W + C \sum_{i=1}^{n} \xi_i
\]

Subject to the constraints:

\[
y_i (W \cdot \varnothing (x_i) + b ) \geq 1 - \xi_i, \text{ and } \xi_i \geq 0, i = 1, \ldots, N
\]

where, $C$ is the capacity constraint, $w$ is denoted as the coefficient vector, $b$ is a constant, $\xi$ represents the parameters for handling the non-separable input data, $y \in \{-1, 1\}$ is the class labels, $x_i$ represents the independent variables and Kernel ($\varnothing$) is denoted as a transform data from the input (independent data) to the feature space. Since the greater value of $C$ results in optimal classification obtained and minimizes the errors of misclassification, the value of $C$ was carefully chosen to avoid overfitting (Table 1).

(ii) Regression SVM (also known as epsilon-SVM regression)

The SVM was also applied to regression where the functional dependence of dependent variables of $y_i$ on a set of independent variables of $x$ were estimated (Cristianini et al. 2000; García Nieto et al. 2013). The corresponding equation can be written as (García Nieto et al. 2013):

\[
\frac{1}{2} W^T W + C \sum_{i=1}^{n} \xi_i + C \sum_{i=1}^{n} * \xi_i
\]

subject to;

\[
w^T \varnothing(\varnothing(x_i)) + b - y_i \leq \epsilon + \xi_i
\]

\[
y_i - w^T \varnothing(\varnothing(x_i)) \leq \epsilon + \xi_i
\]

$\xi_i$, $\xi_i + \geq 0, i = 1, 2, 3, \ldots, N$

| Kernel Type               | Kernel Parameter |
|---------------------------|------------------|
| Radial Basis Function (RBF) | $\Gamma = 0.010638, C = 10$ (where $C$ is Capacity) |

Table 1 | Standard configuration of SVM kernel-RBF tuning parameter
Phase 3: The Kernel-RBF (Radial Basis Functions) trick was chosen with a similar format as that of the Kernels-based learning methods. This is the most preferred and reliable execution type of Kernels-based learning methods with the case-wise as the selected case. Furthermore, the technique of Kernel-RBF method employs the specific function of the non-linear classification for the center of the subsets where two tuning parameters (γ, α) were added, with α being as the kernel parameter, and γ as regularization parameter. These functions are truly the heart of the statistical inference analysis. The value of γ constant was set to 0.185 and the training cost constant capacity, C was set to 10 for non-linearity classification and regression. The results of the used of the kernel trick acts as the problem-solver to the regression with the separable non-linearly in the input space.

In Kernel-RBF, the hidden layer provides a set of functions forming the arbitrary basis as the decision boundary for the oil spill compounds, which is then converted to the input layer of the input vectors or oil spill compounds (SAS141A, SAS141B, BAS303A, BAS304B ... x_m). Subsequently, these layers are expanded or spanned through the hidden layer of kernel inner product (support vectors (x_i)) to which the linear output layer maps the input vectors (training vectors) to optimal discrimination as the output (y). Similar to SVM method, the Kernel inner product provides the optimal hyperplane construction or similar function of neuron network in ANN in the feature space of output. During the training level, the proper weights and support vectors (independent variables) were properly computed, and only those support vectors with a non-zero value, α co-efficient were considered in the calculation (Haykin 1994). The Lagrange classifier method was required to resolve the optimal weight (w) in high dimensional space based on the following expression (Haykin 1994):

\[
w = \sum_{i=1}^{N} \alpha_i d_i \phi(x_i)
\]

(8)

where, α_i represents the coefficients of Lagrange multiplier.

Phase 4: The kernel functions as the algorithm in diverse dataset pattern assessment through data mapping into the high dimensional feature space. The Kernel-Radial Basis Function algorithm was performed using the following equation (Haykin 1994):

\[
k(x_i, x_j) = exp\left(-\frac{|x_i - x_j|^2}{2\delta^2}\right)
\]

(9)

Subject to the optimal weight solution in Equation (9) (Haykin 1994):

\[
d(W^T x_i + b) \geq 1
\]

where i = 1, ..., N are the training samples, d_i = +1 or d_i = -1 are weight vectors and w, in \(O(w) = \frac{1}{2}(W^T W)\).

Phase 5: The Cross Validation method was selected for predictive model validation through assessing the optimal value of different variables to determine the Mean Error Square (MES), coefficient determination (R^2) of fitness factor and correlation coefficient by using a 10-fold cross validation procedure performed by the cross validation algorithm.

RESULT AND DISCUSSION

The Support Vector Machines (SVMs) have recently become one of the most popular machine learning tools for data mining and pattern recognition to solve problems in data clustering (Vapnik 1995; Singh et al. 2011; Ni & Zhai 2016). The outstanding performance of SVMs in classification and regression in many applications has been proven by Cristianini & Shawe-Taylor. The enabling-element values in large quantities and accurate predictability requires the optimum problem solving mechanisms that are commonly sought by many organizations in achieving market objectivity (Meeus et al. 2005). Support Vector Machines (SVMs) selection model which is used for extrapolation, interpolation and thorough evaluation of data in description that is definitely advantageous in this context, and crucially needed to cross-validated or re-confirm the results obtained from GC-FID and GC-MS. The Support Vector Machines (SVMs) approach remarkably improves the complex patterns in data exploitation (oil spill dataset) obtained from GC-FID and GC-MS into clustering and prediction of large mapped input datasets that are structurally classified. The classification function is known as a hyperplane in the space of input oil spill datasets in richer features with a hyperplane clustering or separating space (Alamdar et al. 2016). In this study, SVM-approaches to the classification-based methods enable characterization of the nature of data input behavior with a decision boundary that is separating the oil spill compounds (unseen variables) into a non-linearity separable high-dimensional hyperplane. This allows the oil spill variables to be separated or classified into one or two classes or categories that have similar homogeneity. In oil spill fingerprinting applications, this SVM approach
separates the datasets into high dimensional clustering-based methods that keep the maximal-margin hyperplane between the classes and enables the generative data distributions. Moreover, the approach efficiency for high dimensional datasets, clustering was proven by Moraes & Faria in 2016 (Moraes & Faria 2016).

**Polycyclic aromatic hydrocarbon characterization in oil spill samples**

PAH is commonly known as a highly toxic substance for the environment and its presence could be detrimental to people and other beings. PAHs in petroleum-oil can exist in both low molecular weight and pyrogenic high molecular weight. The aromatic structures mostly contain paraffinic chain, naphthenic and aromatics rings (one to four rings), where the latter structures are in the form of side by side. However, the concentration of PAHs in this study was the lowest concentration detected in the lube oil, but the highest concentration found in the diesel-based spill. In general, the used lube oil contain both low and high molecular weight PAHs resulting from the incomplete combustion of fuel oil and residues (Bishop 1995; Kaplan et al. 2001; Matar & Hatch 2001).

Table 2 presents the summarized quantitation results of 11 selected oil spill samples of PAHs. From the overall quantitation of PAHs in diesel from the oil sample PP5_liq the highest concentrations of $\sum$ Alkylated PAHs (2,221.17 $\mu$g/ml) and $\sum$ EPA PAHs (10.48 $\mu$g/ml) were detected. The diesel from the oil sample GPe_Jetty was second highest with the concentration of $\sum$ Alkylated PAHs (715.04 $\mu$g/ml) and $\sum$ EPA PAHs (11.2 $\mu$g/ml). From the overall 47 oil samples, only 11 were found to be significant and were selected based on the significant detectable concentrations for further elaboration in this section.

The values of the alkylated PAH and the EPA Priority PAH can determine the category of oil spill such as diesel, waste oil (from the mixture of many kinds of oils), lube oil and Bunker C. The alkylated homologue PAHs comprises hydrocarbon compounds of napthanic (C1-N – C4-N), phenanthrene (C0-P – C3-P), dibenzothiophene (C0-D – C4-D), Fluorene (C0-F – C3-F), Chrysene (C0-C – C3-C), while the EPA Priority PAHs consist of acenaphthylene (AcI), acenaphthene (Ace), anthracene (An), Fluoranthene (Fl), Pyrene (Py), Benz[a]anthracene (BaA), Benzo[b]fluoranthene (Bbf), Benzo[k]fluoranthene (Bkf) and Benzo[e]pyrene (Bep). The quantified alkylated homologous PAHs demonstrates, the waste oil that comprises the mixture of many kinds of oils from the oil samples and the suspected Bunker C with the lowest concentration of PAHs, at Kukup (PKK1), Port Klang (KLG-Net), Kuala Kedah (KK1), Prai Port (PP4) and the probable Bunker C with the quantitated values of 3.07 $\mu$g/ml, 3.97 $\mu$g/ml, 0.47 $\mu$g/ml, 0.43 $\mu$g/ml and 1.22 $\mu$g/ml, respectively.

In this study, on the other hand, diesel has revealed the highest concentration of aromatic hydrocarbon ($\sum$ alkylated PAHs) from the oil samples; Pasir Gudang (PP5-liq), Gelang Patah (GPe-Jetty) and Kuala Perlis (KP3) with the quantitated values of 2,221.17 $\mu$g/ml, 715.04 $\mu$g/ml and 67.33 $\mu$g/ml, respectively. This large $\sum$ alkylated PAHs contribution could have resulted from the combustion from marine engines (Yang et al. 2016). Lube oil from Tanjung Gemok (TGG 03) was determined as the second highest, but very low in concentration of total $\sum$ alkylated PAHs with the quantitated value of 3.254 $\mu$g/ml. In general, the lube oil mentioned above is not only very low in terms of concentration of GC-detectable alkanes fraction but also in terms of concentration of the aromatic hydrocarbons (Kaplan et al. 2001; Lu & Kaplan 2008; Yang et al. 2015). Diesel from the oil sample of Pasir Gudang (PP5-liq) showed a significantly high amount of 2-to-6 rings alkylated homologous aromatic hydrocarbon (PAHs) content, in particular the compounds of naphthalene, phenanthrene, dibenzothiophene and fluorene with concentrations that were determined to be 1,185.197 $\mu$g/ml, 408.33 $\mu$g/ml, 101.59 $\mu$g/ml and 3.88 $\mu$g/ml, respectively. However, a relatively low concentration of the US EPA Priority PAHs were found in the oil sample diesel (PP5-liq) with the quantitated value of 10.48 $\mu$g/ml that was dominated by Fluoranthene (1.92 $\mu$g/ml) and pyrene (8.56 $\mu$g/ml).

The reason for the relatively low concentration of the US EPA Priority PAHs in the diesel oil sample might be the presence of the trace amount of polycyclic sulfur aromatic hydrocarbon (PSAH), such as dibenzothiophene, in diesel, instead of alkylated PAHs. Compounds such as benzo[b]naphtol[1-2,d]thiophene are commonly used to trace the existence of diesel-based oil emission in the environment (Daisey et al. 1986). Meanwhile, the lube oil from the TGG03 (Tanjung Gemok, Rompin) is exactly the opposite of diesel in terms of the amount of detectable alkylated PAH and US EPA PAH in this sample, which is $3.25 \mu$g/ml in the former and 0.14 $\mu$g/ml in the latter. The suspected Bunker C from the oil sample MERSING 4 (Mersing, Johor) was the least detectable in terms of concentrations of alkylated PAH and US EPA Priority PAHs, with the quantitated values of 1.22 $\mu$g/ml and 0.01 $\mu$g/ml, respectively. In general the lube oil has very
Table 2 | Target PAH quantitation results from eleven (11) selected oil spill samples

| Compound | Waste oil | Diesel | Mixed oil | Bunker C | Lube oil |
|----------|-----------|--------|----------|----------|---------|
|          | Pasir Gudang (PP2_MAMPU) | Pasir Gudang (PP3-liq) | Gelang patah (GP4-Jetty) | Kuala Perlis (KPS) | Kuala Lumpur (TL 03b) | Port Klang (KLG-Net) | Kuala Kedah (KK1) | Port Prai (PP4) | Mersing (MERSING4) | Tanjung Gemok (TGG 03) |
| Alkylated PAHs | C0-N | 0.21 | 2.14 | 3.64 | 0.09 | 0.10 | 0.18 | 0.49 | 0.08 | 0.09 | 0.08 | 0.04 |
|          | C1-N | 0.65 | 6.55 | 44.18 | 0.11 | 0.12 | 0.28 | 0.54 | 0.10 | 0.11 | 0.54 | 0.05 |
|          | C2-N | 0.32 | 133.21 | 145.21 | 0.07 | 0.09 | 0.33 | 0.45 | 0.07 | 0.07 | 0.07 | 0.07 |
|          | C3-N | 0.23 | 537.97 | 178.69 | 0.08 | 0.57 | 0.26 | 0.34 | 0.08 | 0.05 | 0.06 | 0.36 |
|          | C4-N | 0.38 | 506.10 | 120.58 | 0.13 | 2.15 | 0.33 | 0.54 | 0.02 | 0.02 | 0.03 | 0.25 |
| Σ Naphthalene | 1.79 | 1,185.97 | 492.30 | 0.47 | 3.02 | 1.58 | 2.56 | 0.34 | 0.33 | 0.78 | 0.77 |
|          | C0-P | 0.07 | 16.52 | 5.00 | 0.00 | 0.01 | 0.27 | 0.33 | 0.00 | 0.05 | 0.00 | 0.02 |
|          | C1-P | 0.09 | 92.78 | 13.32 | 0.15 | 0.34 | 0.23 | 0.25 | 0.04 | 0.01 | 0.04 | 0.21 |
|          | C2-P | 0.15 | 150.27 | 32.78 | 0.80 | 1.80 | 0.18 | 0.10 | 0.01 | 0.01 | 0.07 | 0.28 |
|          | C3-P | 0.08 | 102.22 | 19.37 | 1.09 | 1.46 | 0.13 | 0.04 | 0.03 | 0.00 | 0.08 | 0.21 |
|          | C4-P | 0.00 | 46.54 | 27.40 | 0.62 | 0.68 | 0.08 | 0.02 | 0.05 | 0.00 | 0.07 | 0.14 |
| Σ Phenanthrene | 0.38 | 408.33 | 97.87 | 2.66 | 4.28 | 0.89 | 0.74 | 0.15 | 0.08 | 0.26 | 0.87 |
|          | C0-D | 0.01 | 3.48 | 5.34 | 0.00 | 0.01 | 0.03 | 0.04 | 0.00 | 0.00 | 0.00 | 0.39 |
|          | C1-D | 0.01 | 13.13 | 7.56 | 0.02 | 0.04 | 0.04 | 0.05 | 0.00 | 0.00 | 0.00 | 0.09 |
|          | C2-D | 0.00 | 35.26 | 7.56 | 0.20 | 0.42 | 0.05 | 0.05 | 0.00 | 0.00 | 0.02 | 0.19 |
|          | C3-D | 0.00 | 30.24 | 7.35 | 0.44 | 0.46 | 0.05 | 0.03 | 0.00 | 0.00 | 0.03 | 0.17 |
|          | C4-D | 0.00 | 19.48 | 8.86 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Σ Dibenzothiophene | 0.02 | 101.59 | 36.67 | 0.66 | 0.92 | 0.17 | 0.17 | 0.00 | 0.00 | 0.05 | 0.83 |
|          | C0-F | 0.03 | 8.93 | 3.34 | 0.00 | 0.01 | 0.09 | 0.13 | 0.00 | 0.00 | 0.00 | 0.01 |
|          | C1-F | 0.08 | 102.78 | 15.23 | 0.03 | 0.26 | 0.13 | 0.17 | 0.00 | 0.00 | 0.01 | 0.16 |
|          | C2-F | 0.11 | 204.51 | 29.43 | 0.27 | 1.49 | 0.18 | 0.21 | 0.00 | 0.00 | 0.04 | 0.31 |
| Alkylated PAHs | C3-F | 0.00 | 203.52 | 31.71 | 0.89 | 1.97 | 0.23 | 0.16 | 0.00 | 0.00 | 0.08 | 0.28 |
| Σ Flourene | 0.22 | 519.74 | 79.71 | 1.19 | 3.72 | 0.63 | 0.66 | 0.00 | 0.00 | 0.13 | 0.75 |
|          | C0-C | 0.03 | 1.67 | 3.69 | 0.02 | 0.04 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 |
|          | C1-C | 0.00 | 2.45 | 3.05 | 0.02 | 0.05 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.01 |
|          | C2-C | 0.00 | 1.43 | 1.03 | 0.02 | 0.05 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.01 |
|          | C3-C | 0.00 | 0.00 | 0.73 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Σ Chrysene | 0.03 | 3.88 | 4.81 | 0.04 | 0.15 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.02 |
| EPA Priority PAHs | Biphenyl (Bp) | 0.00 | 0.00 | 0.95 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
|          | Acenaphthylene (AcI) | 0.00 | 0.00 | 0.10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 |
|          | Acenaphthene (Ace) | 0.01 | 0.00 | 0.50 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.00 | 0.00 | 0.00 |
|          | Anthracene (An) | 0.01 | 0.00 | 0.32 | 0.00 | 0.01 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.01 |
|          | Fluoranthen (Fl) | 0.01 | 1.92 | 0.60 | 0.01 | 0.03 | 0.02 | 0.01 | 0.00 | 0.01 | 0.00 | 0.02 |
|          | Pyrene (Py) | 0.01 | 8.56 | 2.20 | 0.02 | 0.06 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 | 0.04 |
|          | Benzo[a]anthracene (BaA) | 0.00 | 0.00 | 0.97 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
|          | Benzo(b)fluoranthene (BbF) | 0.00 | 0.00 | 4.90 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

(continued)
unique characteristics as it contains a low concentration of polyaromatic hydrocarbon as well as the alkanes distribution (Kaplan et al. 2001; Yang et al. 2015, 2016).

Based on Table 2, the lube oil (used) has a similar characteristic to the diesel-based oil; however, the concentration of the EPA Priority PAHs is significantly distinctive as the additional constituents of pyrogenic PAHs. For instance, there are two constituents of EPA Priority PAHs detected in the diesel sample of Pasir Gudang (PPS-liq), namely Anthracene (An) and Fluoranthene (Fl) with the concentrations accounting for 1.92 μg/ml and 8.56 μg/ml, respectively. The used lube oil has significantly retained the pyrogenic PAHs ranging from Acenaphthylene (Acl), Acenaphthene (Ace), Anthracene (An), Fluoranthene (Fl), Pyrene (Py), Benz[a]anthracene (BaA), BpH, Perylene (Pe), Benz[e]pyrene (BeP), Biphenyl (Bp), Benz[k]fluoranthene (BkF), Benz[a]pyrene (BaP), Benzo(e)pyrene, Benzo(a)pyrene and Perylene, except for no concentrations of Benz[b]fluoranthene (BbF) and Indeno(1,2,3-cd)pyrene.

The chromatogram signatures provided the information that the used lube oil of TGG 03 could be a mixture of lube oil, unburned diesel, and combustion exhaust of oils from marine use (Yang et al. 2016).

Furthermore, the waste oil from the ambiguous mixture of many types of oil; for example, the oil samples from Port Klang (KLG-Net), Kukup (PKK1), Kuala Kedah (KK1) and Prai Port (PP4) were found to contain both alkylated PAHs and EPA Priority PAHs, all of which were observed in low concentrations (Table 2). Thus, it was determined that the oil mixture-types of many oils could have been blended with petrogenic PAHs and pyrogenic PAHs contributions (from closed system combustions of diesel engines) as the oil samples containing low and high-molecular-weight (HMW) (2–6 ringed) PAHs. However, only oil samples from Port Klang (KLG-Net), Prai Port (PP4) and Kukup (PKK1) were characteristically almost similar by looking at the low and high-molecular-weight alkylated PAHs and EPA Priority PAHs fragmentations. While the other oil samples, like Kuala Kedah (KK1), was only dominant with the alkylated PAHs compounds of naphthalene and phenanthrene, the compounds of Anthracene (An) and Fluoranthene (Fl) dominated the EPA Priority PAHs compound. In addition, through the ambiguous characterization of the suspected Bunker C oil-type (MERSING 4), it probably contained a low concentration of both petrogenic and pyrogenic of PAHs (Table 2).

### Table 2 | continued

| Compound | Waste oil | Diesel | Gelang | Kuala | Tanjung | Mixed oil | Bunker C | Lube oil |
|----------|-----------|--------|--------|--------|---------|-----------|----------|----------|
|          | Pasir Gudang | Pasir Gudang | Gelang | Kuala | Tanjung | Kukup | Port Klang | Kuala | Prai | Mersing | Tanjung |
| Benz[k]fluoranthene (BkF) | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Benzo[a]pyrene (BaP) | 0.00 | 0.00 | 0.07 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Benzo(e)pyrene | 0.07 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Benzo(a)pyrene | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Perylene | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Indeno(1,2,3-cd)pyrene | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Dibenzo(ah)anthracene | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Benzenz(ghi) perylene | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| \( \sum \) Naphthalene | 1.79 | 1,185.97 | 492.30 | 0.47 | 3.02 | 1.38 | 2.36 | 0.34 | 0.33 | 0.78 | 0.77 |
| \( \sum \) Phenanthrene | 0.58 | 408.33 | 97.87 | 2.66 | 4.28 | 0.89 | 0.74 | 0.15 | 0.08 | 0.26 | 0.87 |
| \( \sum \) Dibenzothiophene | 0.22 | 101.59 | 36.67 | 0.66 | 0.92 | 0.17 | 0.17 | 0.00 | 0.00 | 0.05 | 0.83 |
| \( \sum \) Fluorene | 0.22 | 519.74 | 79.71 | 1.19 | 3.72 | 0.63 | 0.66 | 0.00 | 0.00 | 0.15 | 0.75 |
| \( \sum \) Chrysene | 0.00 | 5.88 | 4.81 | 0.05 | 0.15 | 0.01 | 0.03 | 0.00 | 0.01 | 0.00 | 0.02 |
| \( \sum \) Alkylated PAHs | 2.44 | 2,221.17 | 715.04 | 5.04 | 12.12 | 3.07 | 3.97 | 0.47 | 0.43 | 1.22 | 5.25 |
| \( \sum \) EPA Priority PAHs | 0.11 | 10.48 | 11.20 | 0.05 | 0.11 | 0.07 | 0.11 | 0.00 | 0.05 | 0.01 | 0.14 |
| \( \sum \) EPA PAHs/\( \sum \) Alkylated PAHs | 0.05 | 0.00 | 0.02 | 0.01 | 0.01 | 0.02 | 0.03 | 0.01 | 0.07 | 0.01 | 0.04 |
In general, this type of residual oil from the heavy fuel oil was derived from the largely used marine diesel and industrial power generators (Wang & Fingas 2003). The pyrogenic of PAHs contribution was dominated by Anthracene (An), Fluoranthene (Fl), Pyrene (Py) and Benz[a]anthracene (BaA) with the values of 0.00314 μg/ml, 0.00971 μg/ml, 0.002641 μg/ml and 0.00444 μg/ml, respectively. However, the petrogenic PAHs contribution primarily of naphthalene, phenanthrene, dibenzothiophene and fluorene were all dominant in Bunker C. Relatively, the \( \Sigma \)petrogenic PAHs concentration was only 1.22 μg/ml, whereas 0.01 μg/ml accounted for the low amount of EPA Priority PAHs. The small ratio of \( \Sigma \)EPA Priority PAH/\( \Sigma \) Alkylated homologous PAHs, which is 0.01 μg/ml, is a significant indication that the contribution of the most poisonous compound (EPA Priority PAH) in this oil sample is very low. It could be a result of the different chemical composition in the suspected Bunker C type fuel compared to the conventional bunker type fuel. The uncommon profiling shows the preferential loss properties of the bunker-type fuel these days, since it can be a mixture of residual oil and diesel fuel or the lighter oil-types (middle-range distillates) according to the marine use (Wang & Fingas 2003; Stout & Wang 2007).

Figure S1 demonstrates the GC-MS profiling of the eleven (11) selected oil samples. The distribution patterns of PAHs from the GC-MS of the sample (a) PP2 MAMPU, in particular, has achieved the maximum peak at the carbon nC6 as 55.97 μg/ml and 31.13 μg/ml. The wider-resolved hydrocarbon is observed in the wider carbon range of nC9 to nC34 and nC12 to nC30. However, the PAHs concentration is diminished after the carbon nC34 and nC30 from the minutes of 31 and 34, respectively. For (b) PP3-liquid, the carbon range of nC14 – nC30 dominates a small hump, at the minutes of 12 to 26. From the peak distributions of oil-spills, the oil sample (c) GP 6-Jetty dominates a wider-resolved hydrocarbon ranging from nC15 to nC28. The GC-detectable fraction in this sample (a) PKK 1 reveals an unexpected or irregular chemical fingerprint as it is typically attributable to the mixture of many kinds of oils (including the vegetable oils and other petroleum-byproducts that contain complex mixtures of constituents) that enter the seawater in the absence of a certain amount of diesel or any other related hydrocarbon fuel-type.

Identifying the source of oil split as the uncommon chemical fingerprints were obtained might become a problematic process. The highest detectable PAHs concentrations are 12.12 μg/ml at nC19. The highest concentrations in this sample (d) PKK 1 are exhibited at the minutes of 24 and 21, respectively. The oil sample (e) KLG7-net demonstrates the uncommon hydrocarbon fragmentation in the absence of petroleum that is hydrocarbon-oil based. The profiling signature of the hydrocarbon for oil sample (f) KK1 shows almost no existence of alkane’s distribution or polyaromatic hydrocarbon with a very minimal unresolved complete mixture. The absence of the unresolved complete mixture ranges from the carbon nC26 to nC40 at the minutes of 26 to 40. The highest PAHs concentration in this sample is 2.845 μg/ml. For oil sample (g) PP4 the carbon ranges of nC18 to nC19 are the most abundant, with the detected concentrations of 1.073 μg/ml, 1.092 μg/ml, 1.054 μg/ml, 0.739 μg/ml and 1.193 μg/ml at the minute of 18. It is suspected that these PAHs oils are from the petrogenic contribution of marine fuel-oil based, vegetable oil from the restaurants of the ships or diesel-based, which are widely used by many ships at the Prai port. The small hump dominates the (h) PP3, ranging from nC16 to nC28. The highest concentrations are detected as 22.642 μg/ml at nC22 for Mersing 4. This indicates the unique characteristic of chromatograms since they contain waste or used lube oil.

The small hump of a typical unresolved complex mixture (UCM) is observed in (i) MERSING 4 from the broader range of nC29 to nC40 in the absence of alkanes and aromatic hydrocarbon concentrations (Yang et al. 2016) from the minutes of 26 to 40. The oil sample (j) TGG 03 is in dark colour and the chromatogram profiling exhibits a broader range of a large unresolved complex mixture (UCM) in the carbon range of nC26 to nC40 at the minutes of 21 to 34. This profile could have been derived from the mixture of lube oil type product and diesel from the engines of ferries carrying passengers to Tioman Island or boating activities in the area, which are considered under the category of waste oil (WO). The unresolved complex mixture (UCM) is an indication of the pyrogenic contribution (incomplete fuel combustion) and the chemical compositional containing higher molecular weight pyrogenic PAH (Bouloubassi & Saliot 1991; Volkman et al. 1992; Kim et al. 2013; Yang et al. 2016) from the engines of ferries that could causally pollute the area. The oil sample (k) TL4 is very low in PAHs concentrations in the carbon range of nC9 until nC24.

The trained classifier in SVM is non-linear, and all the results obtained from the SVM Kernel-RBF classifier visibly demonstrate the high accuracy of oil spill types’ classification (Figure 5). On the training set of the petroleum-oil compound datasets, the classifier was found to be accurately classified 100% (Vapnik 1995). From the classification output, this analysis highly corresponds to coherent decision boundary and projective subspace estimation. The output of the support
vectors of the training performance as the input of the Kernel-RBF classifier for reliable classification with the values of C (10), γ (0.011) and n (33) are depicted in Table 3.

Table 4 presents the Confusion Matrix of Support Vector Machines presenting the cluster for the overall oil spill samples with Observed (rows) × Predicted (columns). This is the confusion matrix of Kernel-RBF Classification on the training set of the data. The high dimensional nature of this approach leads to the existence of the perfect separability of the oil types’ classification where the computation reveals that the slack variable is $\xi \neq 0$ (Gaonkar et al. 2016), and the C value is considerably high with the confidence level of (0.95) and lower value error (Figure 4).

It can be clearly seen that from 47 oil spill sample data as the input vectors the decision boundary hyperplane of the SVM classifier is determined by 33 significant support vectors (SVs) that have yielded the total of 4 non-linearly separable classes of oil type through the observations (Bunker C – 4 nos, diesel – 15 nos, MO – 22 nos, and WO – 2 nos.). The SVs are the training input vectors subset where the 33 support vectors are the vectors that generate the non-zero Lagrange multiplier coefficients ($\alpha_i$).

Overall, all four (4) nos. bunker C prior cluster were identified based on the chromatogram profiling. Unfortunately, the one identified bunker C in the previous cluster is determined as 100% incorrect. The SVs have predicted the bunker C cluster as non-bunker C product-derivatives, where the SVs classified the bunker C as the MO. Obviously, the total twenty-two (22) nos of MO were 100% correctly classified. The WO has obtained 100% incorrect results as it was predicted by SVs rather than the MO (Figure 5). The lube oil prior cluster; namely, 4 nos, are predicted as diesel-type product. Consequently, the lube oil cluster has not appeared in the table of classification summary of support vector machines (Table 4). The environmental samples that are highly exposed to the adulteration and contamination have altered the physical and chemical characteristics of the WO and lube-oil.

It is rather important to note that the prediction capabilities are captured as the one petroleum-oil spill sample at ‘Dependent’ under the category of MOLFO; however, the SVM classifier predicted the MOLFO as the HFO, whilst the remainder of the oil samples were correctly classified as they fit to the clusters to which they actually belong.

Table 4 | Confusion matrix of support vector machines presenting the cluster for the overall oil spill samples with observed (rows) × predicted (columns)

| Class observed | Bunker C | Diesel | MO | WO |
|----------------|----------|--------|----|----|
| Bunker C       | 0        | 0      | 4  | 0  |
| Diesel         | 0        | 3      | 12 | 0  |
| MO             | 0        | 0      | 22 | 0  |
| WO             | 0        | 2      | 0  | 2  |

Note: Confusion matrix (Support Vector Machines), Cluster, Overall sample (Sheet1 in SVM – PrimaryPrediction Summary.xlsx), SVM: Classification type 1 (C = 4.000), Kernel: Radial Basis Function (gamma = 0.011), Number of support vectors = 33 (3 bounded).

Table 3 | Classification support vector machines summary

| Class name | Total | Correct | Incorrect | Correct (%) | Incorrect (%) |
|------------|-------|---------|-----------|-------------|---------------|
| Bunker C   | 4     | 0       | 4         | 0           | 100           |
| Diesel     | 15    | 3       | 12        | 20          | 80            |
| MO         | 22    | 22      | 0         | 100         | 0             |
| WO         | 2     | 0       | 2         | 0           | 100           |

Figure 4 | An illustration of the hyperplane of the Support Vector Machines (SVM).

Figure 5 | Number of features and classification of oil spill compound with an observed accuracy of 100%.
(Figure 6). The SVM results explain that the Kernel-RBF classifier prediction error by defect was actually based on the actual value according to the oil spill sample background. As we can notice from this analysis, Kernel-RBF has improved the computation time for subspace estimation.

**CONCLUSION**

The study of the seawater PAHs along the Peninsular Malaysia coastlines has revealed the presence of aromatic compound contaminants at most of the sites. Different PAHs concentrations resulted from the spatial variant contaminants from many activities at the sites. The exposure of the oil spills into the environment has led to a high potential of weathering or degradation process, yielding chemically and physically changed contaminants. This study shows that the $\sum$TAH concentration in diesel from the oil sample PP3_liquid has the highest concentrations of $\sum$ Alkylated PAHs (2,221.17 μg/ml) and $\sum$ EPA PAHs (10.48 μg/ml). Diesel from the oil sample GP6_Jetty is the second highest with the concentration of $\sum$ Alkylated PAHs (715.04 μg/ml) and $\sum$ EPA PAHs (11.2 μg/ml). The Classification of SVM Kernel-RBF method in this study is used as a decision boundary approach for classification problems. This method deals with maximizing the classification of the input vectors and maps into the higher dimensional feature space of the inner kernel hidden layer, the error value of which is minimized to achieve the most reliable oil spill type classification. The CSVM Kernel-RBF approaches in oil spill classification, which is in fact a different prediction technique compared to Artificial Neural Networks (ANNs), is responsible for the training process within a short execution time for large quantities of datasets from GC-FID and GC-MS technique fingerprints. The results obtained from this approach are highly promising. It has become evident that the SVM classifier (C-SVM) enables the prediction accuracy of classification achieved (not more than 10% error) from oil spill fingerprinting in Peninsular Malaysia. The confusion matrix output demonstrates the accuracy value in classification of four (4) discriminated oil types, namely, diesel, WO, MO and bunker C. Only diesel and MO are found to be correctly predicted by SVs as the oil types in the environmental samples. The most concerning matters are the kernel-RBF approach provides the oil classification to be optimally performed within a relatively short period of time and faster dataset classifications with slack variables of $\xi \neq 0$ and a coherent decision boundary.

**ACKNOWLEDGEMENTS**

The Authors would like to thank the East Coast Environmental Research Institute (ESERI), UNISZA, Department of Environment (DOE) Malaysia, Department of Chemistry Malaysia, Department of Chemistry, Malaya University, The National University of Malaysia and Integrated Envirotech Sdn. Bhd. for their assistance, data provision and financial support.

**FUNDING**

This study was funded by Integrated Envirotech Sdn. Bhd.

**AVAILABILITY OF DATA AND MATERIAL**

Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

**CODE AVAILABILITY**

The CSVM analysis was performed using the Statistica 13 Dell software 2015.

**AUTHORS’ CONTRIBUTIONS**

All authors contributed to the study conception and design. Material preparation, data collection and analysis were
performed by Azimah Ismail and Hafizan Juahir. Writing, review and editing by Azlina Md. Kassim, Sharifuddin Md. Zain, Hadieh Monajemi, Wan Kamaruzaman Wan Ahmad, Munirah Abdul Zali, Ananthy Retnam, Mohd. Zaki Mohd. Taib, Mazlin Mokhtar and Siti Nor Fazillah Abdullah. Supervision by Saiful Bahri Mohamed, Hafizan Juahir and Mohd Ekhwam Toriman. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available, readers should contact the corresponding author for details.

REFERENCES

Adib, H., Shari, F., Mehranbod, N. & Moradi, N. 2013 Support vector machine based modeling of an industrial natural gas sweetening plant. Journal of Natural Gas Science and Engineering 14, 121–131. https://doi.org/10.1016/j.jngse.2013.06.004.

Ahmad, R. 2009 Tanker with 58,000T naphtha in collision off Malaysia. Reuters, p. Available from: https://www.reuters.com/article/idUSKLR24651.

Alamdar, F., Ghane, S. & Amiri, A. 2016 Neurocomputing on-line twin independent support vector machines. Neurocomputing 1–14. https://doi.org/10.1016/j.neucom.2015.12.062.

Amjadi, N., Daraeepour, A. & Keynia, F. 2010 Day-ahead electricity price forecasting by modified relief algorithm and hybrid neural network. IET Generation, Transmission & Distribution 4, 432–444. https://doi.org/10.1049/iet-gtd.2009.0297.

Bishop, C. M. 1995 Neural Networks for Pattern Recognition. Oxford University Press, Inc, Oxford, UK.

Bouloubassi, I. & Saliti, A. 1991 Sources and transport of hydrocarbons in the Rhone delta sediments (Northwestern Mediterranean). Presented J Anal Chem 20, 765–771.

Cristianini, N., Shawe-Taylor, J. & Shawe-Taylor, D. C. S. R. H. J. 2000 An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Available from: https://books.google.com.my/books?id=PXJn_cvx0AC

Daisey, J. M., Cheney, J. L. & Liow, P. J. 1986 Profiles of organic particulate emissions from air pollution sources: status and needs for receptor apportionment modeling. Journal of the Air Pollution Control Association 36 (1), 17–33. https://doi.org/10.1080/00022470.1986.10466041.

Daling, P. S., Faknness, L. G., Hansen, A. B. & Stout, S. A. 2002 Improved and Standardized Methodology for oil Spill Fingerprinting. Available from: https://inis.iaea.org/search/search.aspx?orig_q=RN:33031417

Dufrenois, F. & Noyer, J. C. 2016 One class proximal support vector machines. Pattern Recognition 52, 96–112. https://doi.org/10.1016/j.patcog.2015.09.036.

Gaganis, V. & Pasadakis, N. 2006 Characterization of oil spills in the environment using parallel factor multway analysis. Analytica Chimica Acta 575–574, 328–332. https://doi.org/10.1016/j.aca.2006.03.071.

Gaonkar, R., Yallappa, S., Dhananjaya, B. L. & Hegde, G. 2016 Development and validation of reverse phase high performance liquid chromatography for citral analysis from essential oils development and validation of reverse phase high performance liquid chromatography for citral analysis from essential oils. Journal of Chromatography B 1056, 50–56. https://doi.org/10.1016/j.jchromb.2016.10.001.

García Nieto, P. J., Alonso Fernández, J. R., De Cos Juez, F. J., Sánchez Lasheras, F. & Díaz Muñiz, C. 2013 Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (Northern Spain). Environmental Research 122, 1–10. https://doi.org/10.1016/j.envres.2013.01.001.

Haykin, S. 1994 Neural Networks; A Comprehensive Foundation, 1st edn. Macmillan Publishing, New York, NY, USA.

Ismail, A., Toriman, M. E. M. E., Juahir, H., Kassim, A. M. A. M., Zain, S. M. S. M., Ahmad, W. K. W. W. K. W. & Yusri, M. A. M. 2016 Chemometric techniques in oil classification from oil spill fingerprinting. Marine Pollution Bulletin 111 (1–2), 339–346. https://doi.org/10.1016/j.marpolbul.2016.06.089.

Ismail, A., Mohamed, S. B., Juahir, H., Toriman, M. E., Kassim, A., Zain, S. M. & Yang, C. 2018 DMAIC six sigma methodology in petroleum hydrocarbon oil classification. International Journal of Engineering & Technology 7, 98–106. https://doi.org/10.14419/ijet.v7i5.14.16686.

Juahir, H., Ismail, A., Mohamed, S. B., Toriman, M. E., Kassim, A. M. Zain, S. M. S. M., Ahmad, W. K. W. W. K. W. & Yusri, M. A. M. 2016 Chemometric techniques in oil classification from oil spill fingerprinting. Marine Pollution Bulletin 111 (1–2), 339–346. https://doi.org/10.1016/j.marpolbul.2016.06.089.

Kaplan, I. R., Lu, S., Forenics, Z., Lane, Z., Obispo, S. L., Alimi, H. M. & Obispo, S. L. 2001 Fingerprinting of high boiling hydrocarbon fuels, asphalts and lubricants. 231–248. https://doi.org/10.1006/enfo.2001.0053.

Khelifa, A., Fieldhouse, B., Wang, Z., Yang, C., Landriault, M., Fingas, M. & Pjontek, D. 2007 A laboratory study on formation of oil-SPM aggregates using the NIST standard reference material 1941b. In Proceedings of the 30th Arctic and Marine Oilspill Program, AMOP Technical Seminar, 1, pp. 35–47.

Kim, M., Hee, S., Won, J., Hyuk, U., Jung, J., Yong, S. & Shim, W. J. 2015 Petroleum hydrocarbon contaminations in the intertidal seawater after the Hebei Spirit oil spill–effect of tidal cycle on the TPH concentrations and the chromatographic characterization of seawater extracts. Water Research 47 (2), 758–768. https://doi.org/10.1016/j.watres.2012.10.050.

Lavine, B. K., Brzozowski, D., Moore, A. J., Davidson, C. E. & Mayfield, H. T. 2001 Genetic algorithm for fuel spill
Lu, S. & Kaplan, I. R. 2006 Characterization of motor lubricating oils and their oil – water partition. Environmental Forensics 9 (4), 295–309. https://doi.org/10.1080/15275920802119441.

Matar, S. & Hatch, L. F. 2001 Chemistry of Petrochemical Processes. Elsevier, Amsterdam, the Netherlands. Available from: https://books.google.com.my/books?id=wp5ZV7BQU1UC

Meeus, L., Purchala, K. & Belmans, R. 2005 Development of the internal electricity market in Europe. Electricity Journal 18 (6), 25–35.

Moraes, L. A. M. & Faria, L. F. 2007 A stochastic programming approach to liquified natural gas planning. Pesquisa Operacional 36, 151–165. https://doi.org/10.1590/S0101-74382016036010151.

Ni, T. & Zhai, J. 2016 A matrix-free smoothing algorithm for large-scale support vector machines. Information Sciences 358–359, 29–43. https://doi.org/10.1016/j.ins.2016.04.010.

Ramsey, E., Meyer, B. M., Rangoonwala, A., Overton, E., Jones, C. E. & Bannister, T. 2014 Oil source-fingerprinting in support of polarimetric radar mapping of Macondo-252 oil in Gulf coast marshes. Marine Pollution Bulletin 89 (1–2), 85–95. https://doi.org/10.1016/j.marpolbul.2014.10.032.

Rodriguez-galiano, V., Sanchez-castillo, M., Chica-olmo, M. & Chica-rivas, M. 2015 Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001.

Silva, T. C., Silva, M. A. d. & Tabak, B. M. 2017 Systemic risk in financial systems: a feedback approach. In Working Paper Series 461. Central Bank of Brazil.

Singh, K. P., Basant, N. & Gupta, S. 2011 Support vector machines in water quality management. Analytica Chimica Acta 703 (2), 152–162. https://doi.org/10.1016/j.aca.2011.07.027.

Stout, S. & Wang, Z. 2007 Chemical fingerprinting of spilled or discharged petroleum – methods and factors affecting petroleum fingerprints in the environment. Oil Spill Environmental Forensics 1–53. https://doi.org/10.1016/B978-012369523-9.50005-7.

Vapnik, V. 1995 The Nature of Statistical Learning Theory. Springer-Verlag New York, NY, USA.

Volkman, J. K., Holdsworth, D. G., Neill, G. P. & Bavor Jr, H. J. 1992 Identification of natural, anthropogenic and petroleum hydrocarbons in aquatic sediments. The Science of the Total Environment 112, 203–219.

Wang, Z. 2005 Fate and identification of spilled oils and petroleum products in the environment by GC-MS and GC-FID. Energy Sources 25 (6), 491–501. https://doi.org/10.1080/00908310390195570.

Wang, Z. & Fingas, M. F. 2005 Development of oil hydrocarbon fingerprinting and identification techniques. Marine Pollution Bulletin 47 (9–12), 423–452. https://doi.org/10.1016/S0025-326X(03)00215-7.

Wang, Z. D., Yang, C., Fingas, M., Hollebone, B., Peng, X., Hansen, A. B. & Christensen, J. H. 2005 Characterization, weathering, and application of sesquiterpenes to source identification of spilled lighter petroleum products. Environ. Sci. Technol. 39, 8700–8707.

Yang, C., Wang, Z., Hollebone, B. P., Brown, C. E., Yang, Z. & Landriault, M. 2015 Chromatographic Fingerprinting Analysis of Crude Oils and Petroleum Products. In: Handbook of Oil Spill Science and Technology, pp. 93–163. https://doi.org/10.1002/9781118989982.ch5

Yang, C., Yang, Z., Zhang, G., Hollebone, B., Landriault, M., Wang, Z. & Brown, C. E. 2016 Characterization and differentiation of chemical fingerprints of virgin and used lubricating oils for identification of contamination or adulteration sources. Fuel 163, 271–281. https://doi.org/10.1016/j.fuel.2015.09.070.

First received 28 September 2020; accepted in revised form 7 January 2021. Available online 1 February 2021

Downloaded from http://iwaponline.com/wst/article-pdf/83/5/1039/857870/wst083051039.pdf by guest