

Predicting Marine Fuels with Unusual Wax Appearance Temperatures Using One-Class Support Vector Machines

Njideka Chima-Amaeshi¹, Chris O'Malley¹ and Mark Willis¹

Received: 17 September 2024 / Accepted: 31 March 2025
© Harbin Engineering University and Springer-Verlag GmbH Germany, part of Springer Nature 2025

Abstract

Accurate and robust detection of wax appearance (a medium- to high-molecular-weight component of crude oil) is crucial for the efficient operation of hydrocarbon transportation. The wax appearance temperature (WAT) is the lowest temperature at which the wax begins to form. When crude oil cools to its WAT, wax crystals precipitate, forming deposits on pipelines as the solubility limit is reached. Therefore, WAT is a crucial quality assurance parameter, especially when dealing with modern fuel oil blends. In this study, we use machine learning via MATLAB's Bioinformatics Toolbox to predict the WAT of marine fuel samples by correlating near-infrared spectral data with laboratory-measured values. The dataset provided by Intertek PLC—a total quality assurance provider of inspection, testing, and certification services—includes industrial data that is imbalanced, with a higher proportion of high-WAT samples compared to low-WAT samples. The objective is to predict marine fuel oil blends with unusually high WAT values ($>35^{\circ}\text{C}$) without relying on time-consuming and irregular laboratory-based measurements. The results demonstrate that the developed model, based on the one-class support vector machine (OCSVM) algorithm, achieved a Recall of 96, accurately predicting 96% of fuel samples with WAT $>35^{\circ}\text{C}$. For standard binary classification, the Recall was 85.7. The trained OCSVM model is expected to facilitate rapid and well-informed decision-making for logistics and storage when choosing fuel oils.

Keywords Marine fuel; One-class support vector machines; Wax appearance temperature; Wax; Machine learning

1 Introduction

Accurate and robust detection of wax appearance is crucial for efficient hydrocarbon transportation in the petroleum industry. Wax is a medium- to high-molecular-weight component of crude oil (Bian et al., 2019). The wax appearance temperature (WAT), also known as the cloud point, is the lowest temperature at which the smallest amount of wax forms (Coutinho and Daridon, 2005). When crude oil (i.e., waxy crude) cools to its WAT, wax crystals precipitate out of solution, forming deposits on the pipeline as the solubility limit is reached. This process poses a notable challenge in hydrocarbon transportation, as wax deposits can negatively affect fuel flow (Ahmed, 2007). The WAT is a critical

quality assurance parameter, particularly when for modern fuel oil blends. These oils are blended to comply with strict International Maritime Organization (Li et al., 2020) regulations, and such blending operation usually produces fuel oil with high paraffinic content and increased risk of wax deposition. When the WAT values are known, vessel operators can safely store fuel oils above this temperature (VPO, 2019). Wax deposition is also a major cause of stuck pigs and increased pressure drops in pipelines and hydrocarbon processing (Mansourpour et al., 2019). Re-establishing flow in pipelines where crude oil or fuel has solidified below the WAT is expensive because it presents a flow assurance challenge (de Oliveira et al., 2012). Therefore, the efficient and safe operation of crude oil and hydrocarbon pipelines relies on the accurate determination of the WAT (Zhao et al., 2022).

Measuring and detecting WAT is a challenging task influenced by procedural sensitivities and uncertainties (Kok et al., 1996). Factors like cooling rates affect results, with faster cooling rates often leading to underestimated WAT values (Japper-Jaafar et al., 2016). Many known procedures detect wax in crude oil by measuring the size or the number of wax crystals. Some studies have also proposed laser-induced voltage technology for WAT detection (Zhang et al., 2022). The detection of light scattered by wax crystals has

Article Highlights

- Marine fuels with unusual wax appearance temperatures are predicted.
- One class support vector machine is used as the anomaly detection algorithm.
- A high True Positive Rate (TPR) is achieved.

✉ Njideka Chima-Amaeshi
n.chima-amaeshi2@newcastle.ac.uk

¹ School of Engineering, Newcastle University, Newcastle Upon Tyne NE1 7RU, UK

been used to measure WAT values in very low sulfur fuel oils, as specified by ASTM D5773 (Thomas, 2019).

Some methods have determined WAT by microscopically observing the onset of wax crystal formation. However, this approach is questionable, as it relies on a visual assessment of the onset of wax buildup, which might be misleading—it requires a substantial amount of wax buildup to be detectable (Kok et al., 1996). Other techniques have focused on detecting wax at an earlier stage of formation, but they failed to detect the true WAT, particularly under thermal equilibrium conditions. All other known techniques, including the “Cold finger method” and the use of a viscometer to detect the shift from the linear viscosity-temperature relationship of a Newtonian to a non-Newtonian fluid behavior, have proven ineffective for WAT detection (Speight, 2016). Laboratory-based techniques such as differential scanning calorimetry (Kök et al., 2018) and cross-polar microscopy (Taheri-Shakib et al., 2020) are highly time-consuming, causing delays of weeks or months between sample collection and analysis date. These delays may introduce sample disparities (Al Shakhs et al., 2023). Other methods, such as Karl Fischer titrations and near-infrared (NIR) spectroscopy, are also used for WAT measurement (Zhang et al., 2022). The majority of the methods used for WAT detection produce inconsistent results (Japper-Jaafar et al., 2016) and are time-intensive, as many require cooling fuel oil or crude oil at different rates (Uba et al., 2004). In addition, the ASTM D3117 method is limited to specific crude or fuel oil types. This method struggles with darker or opaque oils, as it relies on visual confirmation of wax formation (Al Shakhs et al., 2023).

Although machine learning algorithms have also been successfully applied to predict the WAT of crude oil based on parameters such as density, freezing point, and wax content (Benamara et al., 2020), this study presents the use of the one-class support vector machine (OCSVM) algorithm to predict marine fuel samples with abnormal WAT values. This method will reduce the dependence on time-consuming and uncertain laboratory analytical methods. Notably, this is the first attempt by the data supplier Intertek PLC—a total quality assurance, testing, and certification company—to utilize a machine learning algorithm to predict marine fuel oil samples (rather than their exact WAT values) with unusual WATs. This method provides an opportunity for rapid identification and segregation of high-WAT samples, even with scarce data. This process will be carried out by correlating spectral data with the assay data from laboratory-measured WAT values.

Although ISO 8217 specifies a maximum allowable WAT of $-16\text{ }^{\circ}\text{C}$ for distillate fuel oils to be used in marine vessels (ExxonMobil, 2022), this study focuses on a threshold of $35\text{ }^{\circ}\text{C}$ based on the operational requirements of Intertek PLC and the vessel operators involved. Different crude sources show varying WAT values based on their origin

(Kok et al., 1996).

The real-world industrial NIR spectral data provided by Intertek PLC are highly imbalanced in favor of samples exhibiting higher WAT values ($>35\text{ }^{\circ}\text{C}$) and a lack of data for samples with lower WAT values ($<35\text{ }^{\circ}\text{C}$). This imbalance makes standard binary classification algorithms ineffective, necessitating the use of an anomaly detection algorithm.

This study aims to showcase the OCSVM algorithm and its ability to classify and predict marine fuel samples based on spectral data from an imbalanced dataset (Cohen et al., 2004).

The remainder of the paper will focus on describing the OCSVM algorithm and the methodology used in this study.

2 One-class support vector machine (OCSVM)

The OCSVM algorithm was developed by Schölkopf et al. (1995), and it is an extension of the SVM algorithm. OCSVM is based on the study that showcased the possibility of building an optimal hyperplane in high-dimensional feature space by maximizing the margin between the origin and the hyperplane (Schölkopf et al., 2001) (Figure 1). Where support vectors refer to the data points close to the separating hyperplane. training samples are samples used to train the model, $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel function, the outliers are samples that are dissimilar to the trained class samples. $f < 0$ is the negative hyperplane. $f > 0$ shows the positive hyperplane, while $f = 0$ is the optimal hyperplane. $\frac{p}{\|w\|}$ represents the distance from the origin to the optimal/decision hyperplane (Fletcher, 2009).

OCSVM is designed to detect anomalous behaviors in processes where data availability for such abnormalities is complicated. This explains why OCSVM is used in novelty and fault detection since many processes store data generated with optimal conditions, resulting in data scarcity for suboptimal situations (Xiao et al., 2015).

As a semi supervised technique, there is no requirement for all the data to be labelled, which mitigates the problems experienced with supervised algorithms, making OCSVM attractive in many industries, including in clinical applications (Lang et al., 2020). To detect nosocomial infections in healthcare environments, a true positive rate (TPR) of 92.6% was achieved with OCSVM, in contrast to a TPR of 50.6% achieved with a regular SVM for binary classification. This was a remarkable result because the available data were greatly imbalanced, providing only 11% of abnormal (infected) cases and 89% of normal cases (Cohen et al., 2004). In the processing industry, OCSVM was useful in the detection of security breaches in a water treatment plant, using data collected during normal operations from subprocesses as single-class training data (Yau et al., 2020). In the oil and gas exploration and production industry, OCSVM

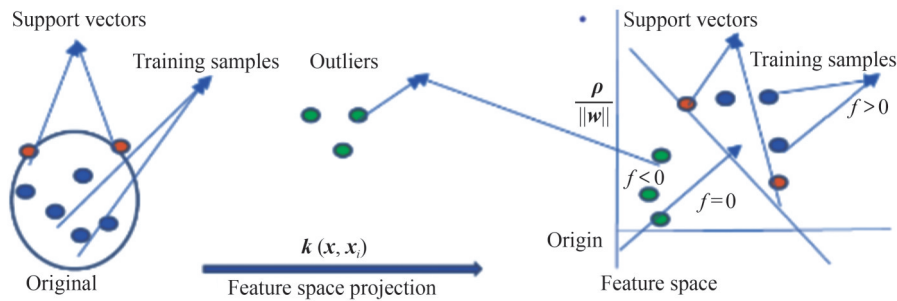


Figure 1 Mapping data onto a higher dimensional space to enable a linear separation in that feature space

was applied in combination with a segmentation algorithm to detect abnormal sensor information in turbomachines in offshore oil installations, enabling effective decision-making by stakeholders. This was seen as a resounding success because of the absence of labelled data generated from the process, which made it difficult to employ a standard supervised learning algorithm (Martí et al., 2015). OCSVM builds and trains the algorithm with the available data from a single class and proceeds to detect samples (or outliers) that are dissimilar to the trained class samples with the aid of a decision boundary. The decision boundary (or hyperplane) separates the outliers from these training samples (Seo, 2007). The hyperplane maximizes the distance from the origin to the samples, with the samples lying far away on the opposite side of the origin.

The margin maximization is achieved by solving the primary optimization (or quadratic programming) problem in the following equation (Seo, 2007), where the origin is positive (Xiao et al., 2015) and a negative label represents the second class as follows:

$$\begin{aligned} \text{minimize } P(\mathbf{w}, \rho, \xi_i) &= \frac{1}{2} \mathbf{w} \mathbf{w}^T + \frac{1}{\nu l} \sum_{i=1}^l \xi_i - \rho \quad (1) \\ \text{s.t. } (\mathbf{w} \cdot \boldsymbol{\phi}(x_i)) &\geq \rho - \xi_i, \xi_i \geq 0, \nu \in [0, 1] \end{aligned}$$

where parameter ν is defined by the user. It explains the upper bound of the training errors or outliers and the lower bound of the fraction of support vectors. This means that it controls the trade-off between maximizing the distance between the hyperplane and the origin and incorporating most of the data points inside the hyperplane region. The offset is defined by ρ , and $\mathbf{w}$ is the weight of the vectors. The transformed image of x_i in the Euclidean space is depicted by $\boldsymbol{\phi}(x_i)$, where $i \in [l]$ and the non-zero slack variable is ξ (Long and Buyukozturk, 2014).

As established earlier, OCSVM aims at classifying data objects based on their similarity to the trained single class. Objects are rejected and classified as outliers if they are dissimilar to the trained class.

For linearly separable data, the hyperplane's distance is $(\mathbf{w} \cdot \boldsymbol{\phi}(x_i)) \geq \rho$ and $\rho > 0$.

The result of the solution of $\min_{\mathbf{w} \in F} \frac{1}{2} \|\mathbf{w}\|$ with $(\mathbf{w} \cdot \boldsymbol{\phi}(x_i)) \geq \rho$ for every i is the distinct hyperplane closer to the origin than all other data points with a maximum distance from the origin to any other hyperplane.

In reality, not all data can be easily divided linearly into positive and negative cases; therefore, a certain error in classification is allowed (Seo, 2007). To solve the linearly inseparable problem, as earlier shown, a kernel is used to map the data onto a higher dimensional space to enable a linear separation in that feature space. Different types of kernels are applicable in OCSVM; however, the Gaussian radial basis function (RBF) is the most commonly used kernel (Cohen et al., 2004), with an ability to handle non-linear problems, leading to a classifier with a universal optimal predictor with lower estimation errors (Martí et al., 2015). This paper uses the Gaussian RBF kernel.

A kernel function $k(\mathbf{x}, \mathbf{y})$ for a feature vector in a low dimension $\mathbf{x}$ and $\mathbf{y}$ is expressed as follows (Long and Buyukozturk, 2014):

$$k(\mathbf{x}, \mathbf{y}) = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{y}) \quad (2)$$

The RBF kernel is expressed as follows (Long and Buyukozturk, 2014):

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right) \quad (3)$$

In Equation (3), σ is the optimizable Gaussian distribution hyperparameter. The model's performance hugely depends on the accurate selection of the kernel parameter (Xiao et al., 2015). A larger value of the Gaussian parameter translates to a narrow distribution with a more concentrated sample distribution. Conversely, a smaller value of the parameter depicts a broader Gaussian distribution with a denser sample distribution (Nie et al., 2022), thereby overfitting the model and significantly reducing the generalization ability of the OCSVM model, whereas a larger value of σ will under-fit the model, which makes it unable to find even common patterns within the data. The other optimizable parameter ν , as earlier stated, determines the number of training samples that can be classified as outliers. A smaller

value means a lower number of false positives (i.e., low WAT samples predicted as high WAT samples), which is a preferred option. Both parameters can be selected or tuned with the aid of a cross-validation technique if samples of the minority class exist (Long and Buyukozturk, 2014).

To classify a new marine fuel sample data point $\mathbf{x}$, Lagrange multipliers α_i and the kernel trick are employed to solve the dual problem as follows (Long and Buyukozturk, 2014):

$$\min_{\alpha} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) \quad (4)$$

$$\text{s.t. } 0 \leq \alpha_i \leq \frac{1}{\nu N}, \quad \sum_i \alpha_i = 1$$

Using the Gaussian equation for mapping onto a feature space the following can be obtained (Long and Buyukozturk, 2014):

$$f(\mathbf{x}) = \text{sgn} \left(\sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) - \rho \right) \quad (5)$$

where ρ can be retrieved as follows (Long and Buyukozturk, 2014):

$$\rho = \sum_j \alpha_j K(\mathbf{x}_j, \mathbf{x}_k), \quad \alpha_j \in \left(0, \frac{1}{\nu N} \right) \quad (6)$$

The support vectors (i.e., all nonzero Lagrangian multipliers α_i) are utilized in solving the decision function for the new sample or data point $\mathbf{x}$ among samples N .

3 Support vector machines

SVM, as proposed by Cortes and Vapnik (1995), was originally intended to be used as a two-class linear classification tool, but it has now been expanded to be applied in nonlinearly separable classification with the use of the kernel trick. The kernel trick is used to map nonlinearly separable data to a higher dimension to enable a class separation (Zhang, 2001; Li et al., 2023), as is the case in this paper. SVMs classify two classes in a dataset with the use of a decision hyperplane. Data objects are plotted as points on a feature space, and each of the features in the data is denoted by a coordinate. To classify the data, the plotted coordinates are divided with a hyperplane to separate the different groups. The coordinates closest to the hyperplane are referred to as the support vectors (Thijssen and Hadjiloucas, 2020). A new and unseen data point is classified and predicted based on the group to which it is closer on either side of the hyperplane (Bangert, 2021).

For linear classification, mathematically, the hyperplane is described as $\mathbf{w}\mathbf{x} + \mathbf{b} = 0$, where $\mathbf{w}$ is normal to the hyper-

plane and $\frac{\mathbf{b}}{\|\mathbf{w}\|}$ describes the distance, which is perpendicular to the hyperplane from the origin (Fletcher, 2009). In Figure 2, if the positive hyperplane is expressed as $\mathbf{w}\mathbf{x} + \mathbf{b} = 1$ and the negative hyperplane is expressed as $\mathbf{w}\mathbf{x} + \mathbf{b} = -1$, the optimal hyperplane is mathematically shown to maximize the margin, which is the perpendicular distance between the hyperplanes closest to the optimal hyperplane and the optimal hyperplane on both sides, assuming that sample data points do not fall in between the positive and negative hyperplanes. A detailed mathematical illustration of the maximization of the margin is shown in the work by Liu (2020a).

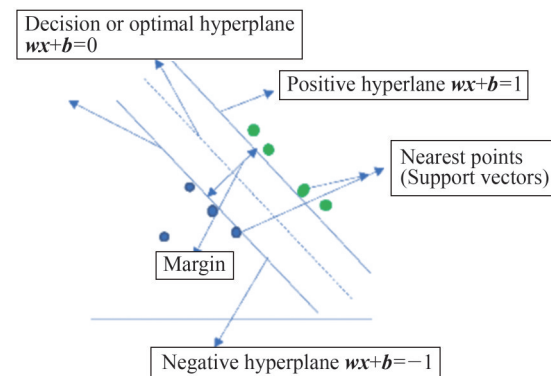


Figure 2 SVM margin and hyperplane

For a sample of marine fuel data with a vector of wave numbers $\mathbf{x}'$, the SVM algorithm is applied for classification and subsequent prediction, as described in the following equation (Liu, 2020b):

$$\mathbf{y}' = \begin{cases} 1, & \text{if } \mathbf{w}\mathbf{x}' + \mathbf{b} > 0 \\ -1, & \text{if } \mathbf{w}\mathbf{x}' + \mathbf{b} < 0 \end{cases} \quad (7)$$

where $\|\mathbf{w}\mathbf{x}' + \mathbf{b}\|$ represents the distance from $\mathbf{x}'$ to the optimal/decision hyperplane. Here, $\mathbf{y}$ is equal to 1 if the distance between the sample and the hyperplane is greater than 0 and -1 if the distance is less than zero and will be assigned to the negative side of the hyperplane, as shown in Figure 2. In the case of marine fuel sample prediction, $\mathbf{y}$ is equal to 1 (i.e., the positive case or high WAT) if the WAT of the fuel is 35 °C or higher and -1 (i.e., the negative case or low WAT) when it is less than 35 °C by weight. A large distance between the data point and the decision boundary translates to higher confidence in the prediction. This is interpreted as a higher prediction certainty because the sample is further away from the decision boundary. Parameter C (Fletcher, 2009) is the key to a decision of how much misclassification is acceptable for a given problem, as follows:

$$\min \frac{1}{2} \|\mathbf{w}\| + C \sum_{i=1}^L \xi_i \quad (8)$$

$$\text{s.t. } \mathbf{y}_i (\mathbf{x}_i \mathbf{w} + \mathbf{b}) - 1 + \xi_i \geq 0, \quad \forall_i$$

It influences the compromise that can be made between the margin size and the penalty variable ξ . An effective SVM classification requires the optimization of parameter C . This involves the careful selection of parameter C to control how the misclassifications should be penalized (Fletcher, 2009).

4 Quality of the model

There is a possibility of errors occurring in the class assignment, which can lead to the misclassification of some of the samples. It is, therefore, important to assess the performance and reliability of a model. A classification error indicates the possibility of a flawed classification when the model is applied to unknown samples in the future (Kuligowski et al., 2016).

Although accuracy has been the common performance metric adopted by many researchers, it does not present a true picture of the classifier's accuracy when dealing with unbalanced data. As an illustration, a high classification or predictive accuracy of 95% for unbalanced data provides no information about the misclassified minority class, but it has only provided a piece of information to show that almost all the samples have been assigned to the majority class with 95% accuracy (Ding, 2011).

Most classification algorithms assume that all datasets are evenly distributed with an almost equal number of samples for every group in the data, but unfortunately, this is far from being true. Most real-life datasets are unbalanced. An unbalanced dataset has a skewed distribution, with one class surpassing the other. The classification of such data has been a major hindrance to many researchers because the majority of classifiers favor the majority class, which is sometimes referred to as the negative class, to the detriment of the minority or positive class (Ding, 2011).

It is common knowledge that the validity of classification models is vital for the subsequent prediction of unknown

samples when introduced into supervised algorithms, including SVM. The samples can be assigned to the appropriate groups or classes, but there is a possibility of misclassification. The performance of the model is measured in terms of TPR or Recall (Westerhuis et al., 2008). The OCSVM algorithm also outputs anomaly scores, which are indicators of samples seen as abnormal to the training samples based on the threshold, in this case, zero. These are seen as anomalies (MathWorks, 2023). In addition to the recall or TPR (Sensitivity), Accuracy and Precision will be used to assess the performance of OCSVM for WAT. It is common to assess OCSVM models based on the percentage of the true anomalous samples detected when compared to all anomalous samples in the dataset (Precision). As shown in the following equation (Nguyen et al., 2023), the model's accuracy compares the predicted and original classes (Ding, 2011):

$$\text{Precision} = \frac{\text{Anomalies Detected}}{\text{Anomalies Reported}} \quad (9)$$

In this study, the focus is on Sensitivity, which is the correct prediction of positives (i.e., samples with WAT values above 35 °C). The TPR represents the probability that samples are correctly identified as part of the high WAT samples (or positive class) used to train the model (Broadhurst and Kell, 2006) and labeled correctly (Bekkar et al., 2013).

The following equation (Sun, 2009) describes the performance metrics as they relate to the sensitivity of the model:

$$\text{TPR (Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (10)$$

5 Methodology

This section focuses on the details of the data and the analysis carried out in this study. The experimental methodology used in this paper is depicted by the flowchart (as shown in Figure 3), describing the data cleaning, preprocessing, and modeling steps used with OCSVM and SVM algorithms.

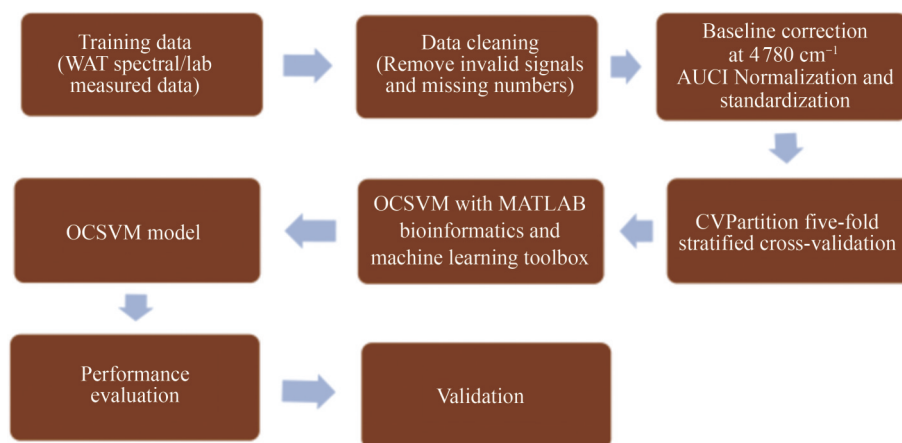


Figure 3 Marine fuel OCSVM and SVM flowchart

5.1 Data collection and analysis

The spectral data were generated by performing 32 scans on each sample of the marine fuel using the ABB MB3000 Series FTIR Spectrometer. This was chosen after several trials, and it has proven to deliver a stable result when performed five times on each sample at a resolution of 4 cm^{-1} to record the absorbance of a typical marine fuel sample covering the wavenumber range of $4\,000\text{--}4\,800\text{ cm}^{-1}$. This range is associated with the fuel's absorption in the combination region range (Blanco and Villarroya, 2002), covering the region for hydrocarbon absorbance (Lammoglia and de Souza Filho, 2011) in line with Intertek PLC's specific operational requirements.

MATLAB Bioinformatics Toolbox was utilized to read and clean the marine fuel spectral data (which are available in Intertek PLC's data repository). Missing values, invalid signals, and redundant values were also removed before computing the average of the five signals to ensure all signals for each sample were represented in the model, bringing the number of valid samples to 368 with 186 wavenumbers.

Figure 4 shows the unbalanced data distribution of the laboratory-measured WAT values, which are used to determine the output variables for this model, which showcases marine fuel samples with high or low WAT values. The distribution reveals 10%–30% of the data (i.e., low WAT samples below $35\text{ }^{\circ}\text{C}$) as the minority class, necessitating a one-class classification, hence the use of OCSVM for anomaly detection. The OCSVM model is designed to detect anomalous behaviors in processes where data availability for such abnormalities poses a challenge (Xiao et al., 2015), as is the case with WAT. More than 80% of the samples belong to the high WAT group based on the threshold of $35\text{ }^{\circ}\text{C}$. In line with the specific operational requirements for Intertek PLC's customer, all marine fuels are stored above $35\text{ }^{\circ}\text{C}$ to prevent wax precipitation and subsequent deposition on the lines.

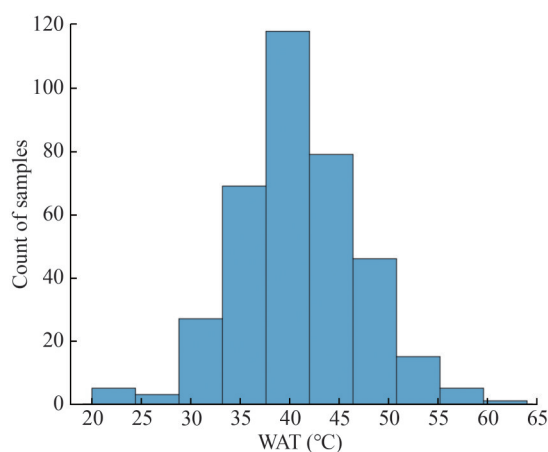


Figure 4 Marine fuel WAT data distribution

5.2 Data preprocessing and model building

MATLAB's Statistics and Machine Learning Toolbox was used to normalize the marine fuel data with the area under the curve set to a unit (Ciaburro and Joshi, 2019) before scaling the same by the standard deviation, resulting in a mean of zero and a standard deviation of 1 (MathWorks, 2022b).

NIR signals suffer from baseline shifts that appear on the absorption axis as a result of additive effects or offsets. These baseline shifts could be a result of several factors, including particle size, porosity, air bubbles, and general morphology attributes. They can also occur as wavelength-based light scattering effects on the sample (Huang et al., 2010), which may appear like absorption but are not related to the chemical composition of the samples and, as such, not relevant in subsequent analyses of the sample (Sandak et al., 2016). These were mitigated using baseline shift techniques.

To ensure the accuracy and robustness of the model, the marine fuel data were normalized, as stated earlier, with the area under the curve set to 1 after carrying out a baseline shift at $4\,780\text{ cm}^{-1}$ to reduce the influence of the signal shifts on the normalized data, restoring all the signal peaks to a common baseline (Fanali et al., 2017). This data was further standardized by subtracting the mean from each sample vector before scaling it by its standard deviation. It is common to normalize the data by setting the area under the curve to 1 in such a way that all the signals are modified and summed to 1. This will create more uniform data with a common baseline and equal contribution to the model (Ciaburro and Joshi, 2019). This practice is recommended for SVM implementation (Meyer et al., 2003) and is in line with Intertek PLC's standard operating procedure.

The OCSVM model was built with MATLAB Statistics and Machine Learning Toolbox (MathWorks, 2022b) for OCSVM using the RBF kernel to implement the quadratic programming algorithm. Hyperparameter tuning was enabled to optimize the nu (ν) parameter. This was carried out with nu (ν) values ranging from 0.1 to 0.8, with 0.5 proving to be the optimum value because it is low enough to prevent overfitting the models and still capture the data complexity (Chen et al., 2001). For internal validation, a five-fold cross-validation (MathWorks, 2022a) technique was implemented in MATLAB. For external validation purposes, 90% (i.e., 274 samples) of the samples with high WAT values were used to train/adjust the model's parameters, whereas the remaining 10% (i.e., 31 samples) was used for validation (or test data) of the model. These test data include 14 low WAT samples, which were used to "contaminate" the model to test its sensitivity. The training data labels were also predicted using the trained classifier as detailed in (MathWorks, 2024).

90% of the data were used because the OCSVM model requires more training samples to compensate for the positive sample instances that may be rejected as a result of

the hard thresholding adopted by the model. Samples are either accepted or rejected as marine fuel samples with an abnormal WAT value as a result of the OCSVM hard thresholding (Guerbai et al., 2014). MATLAB's OCSVM toolbox computes and assigns anomaly scores to each sample according to their similarities to the training samples based on a threshold (i.e., zero). The scores were computed for the training data and compared with the scores for the test data. The scores below 0 are assigned to the samples that are dissimilar to the samples used to train the model and are therefore rejected as anomalies.

For comparison purposes, the same data were also normalized and trained with a standard binary SVM classification algorithm employing the same fivefold cross-validation approach. 90% of the spectral data supplied by Intertek PLC was used to train the model, whereas 10% was used for validation. Similarly, MATLAB Statistics and Machine Learning Toolbox (MathWorks, 2022b) was used to build the model. An RBF kernel scale of 0.002 and a Box Constraint of 0.318 were adopted using Bayesian optimization for hyperparameter tuning, implementing the quadratic programming algorithm for objective function minimization. The "Box Constraint" here is the terminology used in MATLAB to represent the C parameter for the maximization of the margin separation between the support vectors and the hyperplane to minimize the misclassification error of high WAT samples and low WAT samples in the marine fuel dataset (MathWorks, 2022b). A total of 149 support vectors were used to influence the separation of high and low WAT classes.

6 Results and discussion

The proposed OCSVM model was trained using marine fuel samples with WAT values above 35 °C, which accounted for >80% of the total available data. MATLAB assigned anomaly scores to each sample in the training dataset, where positive scores depicted similarity to the training data (WAT >35 °C), whereas negative scores identified outliers (WAT <35 °C).

Figures 5 and 6 show the accuracy and sensitivity plotted against various nu values, respectively, which represent different outlier fractions for each nu value. The results show that accuracy and sensitivity improved as the fraction of outliers detected increased. This observation confirmed that the samples with scores below the threshold of zero were correctly identified as anomalies by the OCSVM classifier and were treated as outliers. Figures 5 and 6 show that the proposed model detected 4% outliers, corresponding to 14 dissimilar samples, though fewer than one sample per test instance prevented clear visualization.

A TPR of 96 indicates that 96% of marine fuel samples with WAT >35 °C were accurately classified. The model

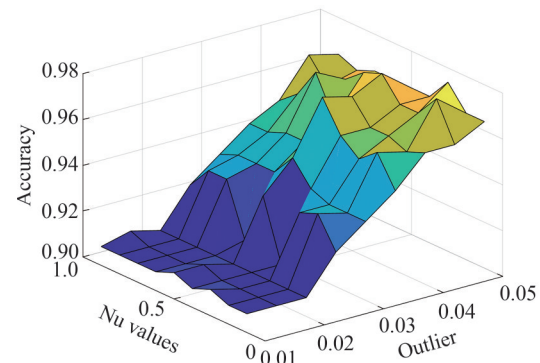


Figure 5 Accuracy, outlier fractions, and nu values

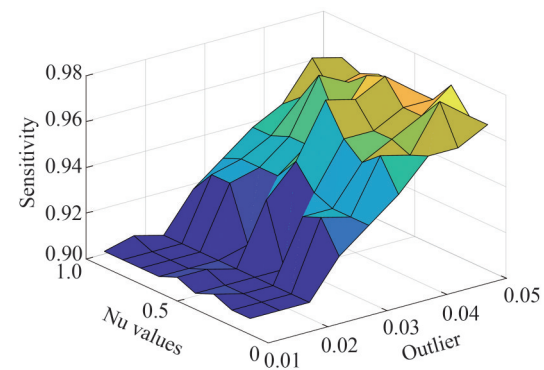


Figure 6 Sensitivity, outlier fractions, and nu values

achieved an accuracy of 0.94 and a precision of 0.99, enabling informed decision-making regarding fuel stability for logistics and storage. These results highlight OCSVM as a high-performing anomaly detection algorithm for marine fuel spectral data. Table 1 presents the original and predicted labels for the test (or validation) data, demonstrating that the model classified all the positive cases. No false negatives (i.e., high-WAT samples misclassified as low-WAT) were recorded.

All fuels predicted to have WAT >35 °C will be stored separately and will not be taken onboard vessels. The results further showcased the validity of OCSVM over standard binary SVM when classifying and predicting with very scarce data. The standard binary SVM model recorded a TPR of 85.7% for the majority class but failed to classify any samples in the minority class. The poor predictive performance of regular two-class SVM was highlighted by a low area under the receiver operating characteristic (ROC) curve (AUROC) of 0.54, as shown in Figure 7. This observation indicates random classification, with ~50% false positives and 50% true positives. This prediction will lead to extremely inefficient decision-making by vessel operators and potential reputational and maintenance costs.

Table 1 shows that the algorithm identified all the samples (high WAT) that are similar to the samples of the class used to train the model. This means that all the positive

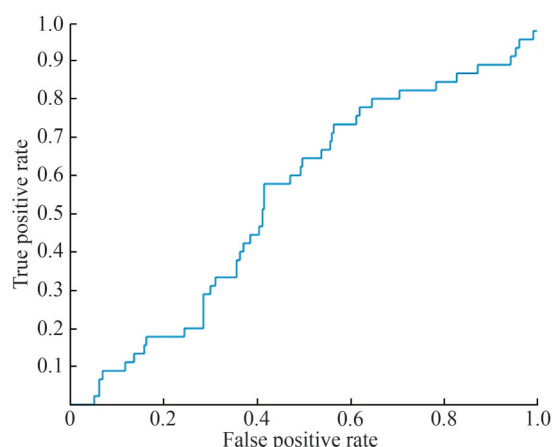


Figure 7 ROC curve of binary SVM on WAT

Table 1 Original data and predicted data class labels

No.	Original class	Predicted class
1	HIGH_WAT	HIGH_WAT
2	HIGH_WAT	HIGH_WAT
3	HIGH_WAT	HIGH_WAT
4	HIGH_WAT	HIGH_WAT
5	HIGH_WAT	HIGH_WAT
6	LOW_WAT	HIGH_WAT
7	LOW_WAT	HIGH_WAT
8	HIGH_WAT	HIGH_WAT
9	HIGH_WAT	HIGH_WAT
10	HIGH_WAT	HIGH_WAT
11	HIGH_WAT	HIGH_WAT
12	HIGH_WAT	HIGH_WAT
13	HIGH_WAT	HIGH_WAT
14	HIGH_WAT	HIGH_WAT
15	HIGH_WAT	HIGH_WAT
16	HIGH_WAT	HIGH_WAT
17	HIGH_WAT	HIGH_WAT
18	HIGH_WAT	HIGH_WAT
19	HIGH_WAT	HIGH_WAT
20	LOW_WAT	HIGH_WAT
21	LOW_WAT	HIGH_WAT
22	LOW_WAT	HIGH_WAT
23	LOW_WAT	HIGH_WAT
24	LOW_WAT	HIGH_WAT
25	LOW_WAT	HIGH_WAT
26	LOW_WAT	HIGH_WAT
27	LOW_WAT	HIGH_WAT
28	LOW_WAT	HIGH_WAT
29	LOW_WAT	HIGH_WAT
30	LOW_WAT	HIGH_WAT
31	LOW_WAT	HIGH_WAT

cases (samples with high WAT values) were successfully predicted by the model. The lower WAT classes were not predicted and there were no false negatives (high WAT samples predicted as low WAT samples) recorded.

7 Conclusions

Herein, we demonstrated the effectiveness of the application of the OCSVM algorithm in predicting suboptimal marine fuel samples using spectral data and laboratory-measured WAT values provided by Intertek PLC. Compared with the standard binary classification algorithm, such as SVM, OCSVM enables the rapid detection of fuel samples with abnormal WAT values, reducing reliance on irregular laboratory test methods.

OCSVM achieved a TPR of 96 with no false negatives using a five-fold stratified cross-validation approach in MATLAB to avoid model overfitting. This capability supports fast and well-informed decision-making for logistics and storage, even when choosing fuel oils on board the ocean-going vessels, despite the absence of sufficient data on LOW-WAT samples required for standard binary classification. In contrast, the regular SVM model yielded a TPR of 85.7, which was confirmed as a chance classification based on the ROC curve. OCSVM classified 4% of samples with WAT of <35 °C as outliers, indicating their dissimilarity from the training data. Although the results are promising, the proposed technique can be improved to detect more LOW-WAT samples (<35 °C). Overall, using OCSVM is recommended, particularly when dealing with data scarcity, as is the case with Intertek PLC's dataset.

Funding Newcastle University and EPSRC (Grant No. 2020/21 DTP: ref. EP/T517914/1).

Acknowledgement The authors would like to thank Newcastle University and EPSRC for the grant 2020/21 DTP: ref. EP/T517914/1 awarded for this project (Realtime Hydrocarbon Composition Measurement) and the technical support and provided. We are also grateful to our partners (Paul Winstone and Luis Tomas) of Intertek Group PLC, Northpoint, Aberdeen Energy Park, Exploration Dr, Bridge of Don, and Aberdeen AB23 8HZ for supplying the data used for this study.

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

Reference

- Ahmed T (2007) 6.2.11.1 Phase Behavior of Waxes. Equations of State and PVT Analysis-Applications for Improved Reservoir Modeling. Houston: Gulf Publishing Company 498
- Al Shakhs MH, Libby C, Chau KJ, Molla S, Sieben VJ (2023) Wax appearance temperature in crude oils measured by surface plasmon resonance. *Petroleum Science and Technology* 43(4): 408-425. <https://doi.org/10.1080/10916466.2023.2293247>
- Bangert P (2021) 3.3.3 Support Vector Machines. *Machine Learning and Data Science in the Oil and Gas Industry-Best Practices, Tools, and Case Studies*, Alpharetta: Elsevier 48-49. <https://doi.org/10.1016/B978-0-12-820714-7.00003-0>
- Benamara C, Gharbi K, Nait Amar M, Hamada B (2020) Prediction

- of wax appearance temperature using artificial intelligent techniques. *Arabian Journal for Science and Engineering* 45(2): 1319-1330. <https://doi.org/10.1007/s13369-019-04290-y>
- Bekkar M, Djemaa HK, Alitouche TA (2013) Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications* 3(10): 28-31
- Bian XQ, Huang JH, Wang Y, Liu YB, Kaushika Kasthuriarachchi DT, Huang LJ (2019) Prediction of wax disappearance temperature by intelligent models. *Energy & Fuels* 33(4): 2934-2949. <https://doi.org/10.1021/acs.energyfuels.8b04286>
- Blanco M, Villarroya I (2002) NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry* 21(4): 240-250. [https://doi.org/10.1016/S0165-9936\(02\)00404-1](https://doi.org/10.1016/S0165-9936(02)00404-1)
- Broadhurst DI, Kell DB (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2(4): 171-196. <https://doi.org/10.1007/s11306-006-0037-z>
- Chen Y, Zhou XS, Huang TS (2001) One-class SVM for learning in image retrieval. *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205)*. Piscataway: IEEE 34-37. <https://doi.org/10.1109/ICIP.2001.958946>
- Ciaburro G, Joshi P (2019) 1.6 Normalization. *Python Machine Learning Cookbook (2nd Edition)* 19
- Cohen G, Hilario M, Sax H, Hugonnet S, Pellegrini C, Geissbuhler A (2004) An application of one-class support vector machines to nosocomial infection detection. *Studies in Health Technology and Informatics* 107(Pt 1): 716-720
- Coutinho JA, Daridon JL (2005) The limitations of the cloud point measurement techniques and the influence of the oil composition on its detection. *Petroleum Science and Technology* 23(9-10): 1113-1128. <https://doi.org/10.1081/LFT-200035541>
- Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3): 273-297. <https://doi.org/10.1007/BF00994018>
- de Oliveira MCK, Teixeira A, Vieira LC, de Carvalho RM, de Carvalho ABM, do Couto BC (2012) Flow assurance study for waxy crude oils. *Energy & Fuels* 26(5): 2688-2695. <https://doi.org/10.1021/ef201407j>
- Ding Z (2011) Diversified ensemble classifiers for highly imbalanced data learning and its application in bioinformatics. *Atlanta: Georgia State University* 1-18
- ExxonMobil (2022) ISO 8217 Marine Fuel Characteristics Definitions. ExxonMobil. Available at: ISO 8217 marine fuel oil characteristic definitions ExxonMobil Marine (Accessed: 16 Sept 2022)
- Fanali S, Haddad PR, Poole CF, Riekkola M-L (2017) 21.3.3 Normalization. *Liquid Chromatography-Fundamentals and Instrumentation, Volume 1 (2nd Edition)* 518-519
- Fletcher T (2009) Support vector machines explained. *Tutorial Paper* 1-19
- Guerbai Y, Chibani Y, Hadjadj B (2014) The effective use of the One-Class SVM classifier for reduced training samples and its application to handwritten signature verification. *2014 international conference on multimedia computing and systems (ICMCS)*. Piscataway: IEEE 362-366
- Huang J, Romero-Torres S, Moshghbar M (2010) Practical considerations in data pre-treatment for NIR and raman spectroscopy. *American Pharmaceutical Review*. <http://www.americanpharmaceuticalreview.com/Featured-Articles/116330-Practical-Considerations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy>
- Japper-Jaafar A, Bhaskoro PT, Mior ZS (2016) A new perspective on the measurements of wax appearance temperature: Comparison between DSC, thermomicroscopy and rheometry and the cooling rate effects. *Journal of Petroleum Science and Engineering* 147: 672-681. <https://doi.org/10.1016/j.petrol.2016.09.041>
- Kok MV, Létoffé JM, Claudy P, Martin D, Garcin M, Volle JL (1996) Comparison of wax appearance temperatures of crude oils by differential scanning calorimetry, thermomicroscopy and viscometry. *Fuel* 75(7): 787-790. [https://doi.org/10.1016/0016-2361\(96\)00046-4](https://doi.org/10.1016/0016-2361(96)00046-4)
- Kök MV, Varfolomeev MA, Nurgaliev DK (2018) Wax appearance temperature (WAT) determinations of different origin crude oils by differential scanning calorimetry. *Journal of Petroleum Science and Engineering* 168: 542-545. <https://doi.org/10.1016/j.petrol.2018.05.045>
- Kuligowski J, Pérez-Guaita D, Quintás G (2016) Application of discriminant analysis and cross-validation on proteomics data. *Statistical Analysis in Proteomics* 1362: 175-184. https://doi.org/10.1007/978-1-4939-3106-4_11
- Lammoglia T, de Souza Filho CR (2011) Spectroscopic characterization of oils yielded from Brazilian offshore basins: Potential applications of remote sensing. *Remote Sensing of Environment* 115(10): 2525-2535. <https://doi.org/10.1016/j.rse.2011.04.038>
- Lang R, Lu R, Zhao C, Qin H, Liu G (2020) Graph-based semi-supervised one class support vector machine for detecting abnormal lung sounds. *Applied Mathematics and Computation* 364: 124487. <https://doi.org/10.1016/j.amc.2019.06.001>
- Li K, Wu M, Gu X, Yuen KF, Xiao Y (2020) Determinants of ship operators' options for compliance with IMO 2020. *Transportation Research Part D: Transport and Environment* 86: 102459
- Li H, Chen H, Li Y, Chen Q, Fan X, Li S, Ma M (2023) Prediction of the optical properties in photonic crystal fiber using support vector machine based on radial basis functions. *Optik* 275: 170603. <https://doi.org/10.1016/j.ijleo.2023.170603>
- Liu Y (2020a) 3.1.2 Scenario 2-Determining the Optimal Hyperplane. *Python Machine Learning by Example (3rd Edition)*. Birmingham: Packt Publishing 78-79
- Liu Y (2020b) 3.1.3 Scenario 3-Handling Outliers. *Python Machine Learning by Example (3rd Edition)*. Birmingham: Packt Publishing 80-81
- Long J, Buyukozturk O (2014) Automated structural damage detection using one-class machine learning. *Dynamics of Civil Structures*. Luemborg: Springer 117-128. <http://hdl.handle.net/1721.1/90062>
- MathWorks (2022a) Cvppartition. Available at: Partition data for cross-validation-MATLAB-MathWorks United Kingdom (Accessed: 07 June 2022)
- MathWorks (2022b) Fitsvm. Available at: Train support vector machine (SVM) classifier for one-class and binary classification-MATLAB fitsvm-MathWorks United Kingdom (Accessed: June 06 2022)
- MathWorks (2023) OneClassSVM. Available at: One-class support vector machine (SVM) for anomaly detection-MATLAB-MathWorks United Kingdom (Accessed: 25 July 2023)
- MathWorks (2024) ResubPredict. Available at: resubPredict-Classify training data using trained classifier-MATLAB-MathWorks United Kingdom
- Martí L, Sanchez-Pi N, Molina JM, Garcia ACB (2015) Anomaly detection based on sensor data in petroleum industry applications. *Sensors* 15(2): 2774-2797. <https://doi.org/10.3390/s150202774>
- Mansourpoor M, Azin R, Osfouri S, Izadpanah AA (2019) Study of wax disappearance temperature using multi-solid thermodynamic model. *Journal of Petroleum Exploration and Production Technology* 9(1): 437-448. <https://doi.org/10.1007/s13202-018-0480-1>
- Meyer D, Leisch F, Hornik K (2003) The support vector machine under test. *Neurocomputing* 55(1-2): 169-186. [https://doi.org/10.1016/S0925-2312\(03\)00431-4](https://doi.org/10.1016/S0925-2312(03)00431-4)

- Nguyen TBT, Liao TL, Vu TA (2023) Anomaly detection using one-class SVM for logs of juniper router devices. arXiv. https://doi.org/10.1007/978-3-030-30149-1_24
- Nie J, Dong Y, Zuo R (2022) Construction land information extraction and expansion analysis of Xiaogan City using one-class support vector machine. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15: 3519–3532. <https://doi.org/10.1109/JSTARS.2022.3170495>
- Sandak J, Sandak A, Meder R (2016) Assessing trees, wood and derived products with near infrared spectroscopy: hints and tips. *Journal of Near Infrared Spectroscopy* 24(6): 485–505. <https://doi.org/10.1255/jnirs.1255>
- Schölkopf B, Burges C, Vapnik V (1995) Extracting support data for a given task. *Proceedings of the 1st International Conference on Knowledge Discovery & Data Mining*. Reston: AIAA 252–257
- Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. *Neural Computation* 13(7): 1443–1471. <https://doi.org/10.1162/089976601750264965>
- Seo K-K (2007) An application of one-class support vector machines in content-based image retrieval. *Expert Systems with Applications* 33(2): 491–498. <https://doi.org/10.1016/j.eswa.2006.05.030>
- Speight JG (2016) 5.4.6 Wax Analysis and Wax Appearance Temperature. *Introduction to Enhanced Recovery Methods for Heavy Oil and Tar Sands* (2nd Edition) 222–224
- Sun DW (2009) 4.3 Evaluation of Classification Performances. *Infrared Spectroscopy for Food Quality Analysis and Control* 92–93. <https://doi.org/10.1016/B978-0-12-374136-3.X0001-6>
- Taheri-Shakib J, Shekarifard A, Kazemzadeh E, Naderi H, Rajabi-Kochi M (2020) Characterization of the wax precipitation in Iranian crude oil based on wax appearance temperature (WAT): The influence of ultrasonic waves. *Journal of Molecular Structure* 1202: 127239. <https://doi.org/10.1016/j.molstruc.2019.127239>
- Thomas B (2019) An automatic test method for wax appearance temperatures of VLSFOs. Available at: Riviera-Whitepapers-An automatic test method for wax appearance temperature of VLSFOS (rivieramm.com) (Accessed: 24 October 2022)
- Thijssen P, Hadjiloucas S (2020) 12.3.2 Advances in Support Vector Machine Classifiers. *State Estimation in Chemometrics-The Kalman Filter and Beyond* (2nd Edition) 237. <https://doi.org/10.1016/C2017-0-02894-X>
- Uba E, Ikeji K, Onyekonwu M (2004) Measurement of wax appearance temperature of an offshore live crude oil using laboratory light transmission method. *Nigeria Annual International Conference and Exhibition*. Abuja Paper Number: SPE-88963-MS. <https://doi.org/10.2118/88963-MS>
- VPO (2019) VPS launches automatic test method for wax appearance temperature of VLSFOs. Available at: <https://vpoglobal.com/2019/10/18/vps-launches-automatic-test-method-for-wax-appearance-temperature-of-vlsfos/> (Accessed: 24 October 2022)
- Westerhuis JA, Hoefsloot HC, Smit S, Vis DJ, Smilde AK, van Velzen EJ, van Duijnhoven JP, van Dorsten FA (2008) Assessment of PLS-DA cross validation. *Metabolomics* 4(1): 81–89. <https://doi.org/10.1007/s11306-007-0099-6>
- Xiao Y, Wang H, Xu W (2015) Parameter selection of Gaussian Kernel for one-class SVM. *IEEE Transactions on Cybernetics* 45(5): 941–953. <https://doi.org/10.1109/TCYB.2014.2340433>
- Yau K, Chow KP, Yiu SM (2020) Detecting attacks on a water treatment system using one-class support vector machines. *IFIP Advances in Information and Communication Technology*. 16th Annual IFIP WG 11.9 International Conference on Digital Forensics, 2020. New Delhi: Springer Science and Business Media Deutschland GmbH 95–108
- Zhang S, Sun X, Liu C, Zhang H, Miao X, Zhao K (2022) Characterization of wax appearance temperature of model oils using laser-induced voltage. *Physics of Fluids* 34(6): 067123. <https://doi.org/10.1063/5.0098727>
- Zhang T (2001) An introduction to support vector machines and other kernel-based learning methods. *AI Magazine* 22(2): 103–103
- Zhao K, Li C, Xia X, Fang K, Yao B, Yang F (2022) Optical techniques for determining wax appearance temperature of waxy crude oil. *2021 International Conference on Optical Instruments and Technology: Optoelectronic Measurement Technology and Systems*. Bellingham: SPIE 492–500. <https://doi.org/10.1117/12.2620320>