Journal of Marine Science and Application (2026) 25: 617-629
https://doi.org/10.1007/s11804-025-00674-9

RESEARCH ARTICLE

Predicting Marine Fuel with High Sulphur Content Using Machine Learning Algorithms

Njideka Chima-Amaeshi[1], Chris O'Malley[1] and Mark Willis[1]

Received: 27 December 2024 / Accepted: 22 April 2025

Abstract

Marine transportation is a significant source of air pollution especially around coastal areas with maritime vessels creating 12% of global sulphur oxides emission in 2014 alone. In compliance with International Maritime Organisation (IMO) regulations, the determination of sulphur content of marine fuels is typically carried out using lengthy laboratory-based analyses. The regulations prohibit the use of High-Sulphur Fuel Oil (HSFO) (>0.5% by weight of Sulphur) in Emission Control Areas (ECA). There is a need for a more efficient means of predicting Sulphur content and differentiating between HSFO and Very Low Sulphur Fuel Oil (VLSFO) samples. This study compares the application of a Support Vector Machine (SVM) and Agglomerative Hierarchical Clustering (AHC) algorithm enhanced with Principal Component Analysis for dimensionality reduction purposes to predict HSFO and VLSFO marine fuel samples based on near-infrared (NIR) industrial data from North Sea operations correlated with laboratory-measured sulphur values instead of relying on lengthy laboratory-based measurements. The study also compares the effect of normalising the data by setting the area under the curve to one and standardising it by subtracting the mean of predictor variables and scaling by standard deviation. The results show that although >70% of HSFO samples were accurately predicted with the SVM, a better result was achieved using the unsupervised learning approach of AHC/PCA with >80% of HSFO samples correctly predicted despite the imbalance in the industrial data providing an effective model for the rapid and well-informed decision-making tool for vessel operators. Normalising the area under the curve to one produced similar results to using standardised data.

Keywords Machine learning; Marine fuel; Support vector machines; Agglomerative hierarchical clustering; High sulphur fuel oil; Very low sulphur fuel oil

1 Introduction

A significant amount of air pollution around coastal areas is generated in the marine industry by maritime vessels (Corbett et al., 2007). In 2014, these vessels were responsible for 12% of global Sulphur oxides emission because of fuelling these vessels (Gu et al., 2025) and currently between 4–9% of global Sulphur oxides emissions (Fan et al., 2023). One study pointed out the emission of more than 1 600 tons of Sulphur oxides with a total environmental pollution cost of more than 40 million euros from shipping in Bandirma port in Turkey (Kuzu et al., 2021). Particulate matter and Sulphur dioxide from marine fuels used in powering sea transportation are some of the major culprits of air pollution (Eyring et al., 2010). Shipping is also a major cause of ocean acidification due to the presence of Sulphur compounds. These are byproducts of the burning of Sulphur-containing fuel oils (Hassellöv et al., 2013). These pollutants are also associated with cardiopulmonary and lung cancer in humans (Corbett et al., 2007).

Vessel operators have resorted to installing scrubbers and/or switching between fuels when in and outside the ECAs. These have proven to be inefficient in terms of cost implications (Gu et al., 2025). Therefore, to efficiently comply with more stringent emissions regulations in the marine industry as specified by the International Maritime Organisation (IMO) on the marine fuel refining processes, shipping emissions, and marine fuel mix (Van et al., 2019), all marine bunker fuels are analyzed according to International Standard Organisation (ISO) 8217 specifications. The regulation is specifically taken from Annex VI International

Article Highlights

- Marine fuel samples with high and low sulphur content are predicted.
- Support Vector Machines and Agglomerative Hierarchical Clustering algorithms were applied.
- The Agglomerative Hierarchical Clustering algorithm achieved a higher True Positive Rate.
- Normalisation methods had an insignificant effect on both algorithms' performance.

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

[1] School of Engineering, Newcastle University. Newcastle Upon Tyne. NE1 7RU

Convention for the Prevention of Pollution from Ships: The Marine Pollution Convention (MARPOL). These specifications provide an insight into the quality of marine fuels to the buyers and allow them to comply fully with the safety and environmental regulations.

IMO regulations which are expected to reduce maritime-related Sulphur emissions from vessels sailing in international waters by 77% (Ju and Jeon, 2022), stipulate that the use of high-Sulphur fuel oil is prohibited in vessels operating in Emission Control Areas (ECA) e.g. Baltic Sea, North Sea and the English Channel (Cullinane and Bergqvist, 2014). This has continued to pose a challenge to the marine industry as they continue to see a high demand for lower Sulphur fuel oil. Based on this, approximately 2 million barrels of oil per day were required to be converted to distillates to enable the refineries to meet the growing demand for low-Sulphur fuel oil in 2020. This is expected to cause a reduction in the demand for residual fuel oil by 150 000 bbl. per day (Concawe, 2016). The sulphur content of marine fuel was restricted to 0.1% (wt.%) in Sulphur Emission Control Areas (SECAs) and Nitrogen Oxide Emission Control Areas (NECAs) in 2015. In 2019, a 0.1% limit on Sulphur emission was established within the Sulphur Emission Control Areas and 3.5% outside these areas (Zis and Cullinane, 2020) According to IMO, this will significantly reduce the adverse effects of shipping emissions on human health and the environment (Van et al., 2019). A review by IMO in 2020 led to the implementation of a regulation with a requirement for a 0.5% Sulphur content by weight limit in marine fuel oil for maritime vessels operating globally (Bilgili, 2021).

It is therefore crucial that the Sulphur content of all marine fuel oil be determined promptly as opposed to the use of time-consuming laboratory-based ASTM methods according to ISO 8217 standards. There are a few ASTM procedures employed in the laboratories to measure Sulphur content in hydrocarbons, some of them involve a general high-pressure decomposition (ASTM D129), and others include the lamp method (ASTM D1266), wavelength dispersive *X*-ray fluorescence Spectroscopy (WDXRF) (ASTM 2622) among others. These methods albeit robust, for example measuring between 1 – 10 parts per million of Sulphur do not generate prompt or instantaneous results (Christopher et al., 2001). According to IHM Marine Surveys, there could be up to 6 days of turn-around time for analyses of fuel oils (IHMMarineSurveys, 2020). Stratiev et al. (2010), reported that detailed laboratory analysis of a crude oil samples can extend into weeks with the cost running into tens of thousands of dollars (Stratiev et al., 2010) necessitating the search for a faster and more accurate software-enabled prediction of the properties for an effective and prompt decision making exercise. For fuels and other hydrocarbon analysis, Near Infrared (NIR) spectroscopy measures absorption by functional groups such as methyl, methylene, aromatic stretching, and others instead of measuring the groups such as naphthenes etc. NIR features are also predominantly showcased in carbonyl-linked C-H stretching, and methoxy C-H stretching as they are mostly dominant in the near-infrared region. These absorption measurements are subsequently correlated with already-known properties in this case Sulphur with the aid of other multivariate analytical tools to determine the properties of unknown materials (Blanco and Villarroya, 2002), (Workman, 2001).

Intertek PLC is a total quality assurance provider of inspection, testing, and certification services to a wide range of industries including Oil and Gas Exploration and Production. For the first time, Intertek PLC is investigating and comparing the application of a classification (Support Vector Machine (SVM) and clustering algorithm (Agglomerative Hierarchical Clustering (AHC)/Principal Component Analysis (PCA) developed and adapted to predict marine fuel samples at different Sulphur levels based on their historic industrial spectral data. The spectral data is correlated with the laboratory-measured Sulphur content by weight (%). The classes identified are High Sulphur Fuel Oil (HSFO) and Very Low Sulphur Fuel Oil (VLSFO). The aim is to enable a swift and informed decision-making exercise to ensure compliance with the strict IMO regulations (Ju and Jeon, 2022) instead of relying on lengthy laboratory measurements. Therefore, this study will elucidate the significance of the application of machine learning algorithms (SVM and AHC) to the clustering, classification and identification of High and Low Sulphur Fuel samples despite the skewness of the data. SVM has been widely investigated and recognized as an excellent classification technique with notable attributes and accuracy. The ability to classify non-linearly separable and high-dimensional data makes SVM one of the most popular classification algorithms in the world of machine learning (Deng et al., 2015). Another strong quality of SVM is the ability to generalize with a low overfitting risk based on the implementation of the Structural Risk Minimization (SRM) as detailed in (Cortes and Vapnik, 1995). SVM has been successfully applied in face recognition (Dadi and Pillutla, 2016), and text categorization (Kapoutsis et al., 2004) among others. Within the hydrocarbon processing industry, the prediction of derived cetane number and carbon/hydrogen ratio in hydrocarbon mixtures from infrared spectra data was possible with SVM (Al Ibrahim and Farooq, 2021). In conjunction with various machine learning algorithms, SVM was helpful in the prediction of the density and viscosity of biofuels (Saldana et al., 2012). The models were subsequently used to estimate densities and dynamic viscosities for samples when laboratory data were not available.

There is no record of existing research on the prediction of high and low-Sulphur marine fuel oil with the aid of Near Infrared Spectroscopy (NIR) coupled with Support Vector Machines or Agglomerative Hierarchical Clustering

algorithm using historic industrial data. Therefore, an efficient algorithm is a necessity to ensure that reliable marine fuel predictions can be made promptly to enable the vessel operators to remain in full compliance with IMO regulations.

This has necessitated this study to investigate and showcase SVM (classification) and AHC (clustering) algorithms' ability to predict and distinguish between marine fuel oil samples with high and low Sulphur content without the need for costly and time-consuming laboratory analysis. This will in turn enable Intertek PLC to provide reliable Sulphur content information to their customers to affect a rapid and more informed decision-making process and ensure compliance with strict marine industry regulations. The clustering algorithm (AHC) will be paired with Principal Component Analysis (PCA) for dimensionality reduction of the features as AHC struggles with large datasets especially when the Ward Linkage technique is used (Lantz, 2023). The supervised algorithm (SVM) relies on the computed support vectors to classify and predict the required outcomes (Hearst et al., 1998). Since this paper is focused on the application of classification and clustering algorithms which have never been utilised by Intertek PLC, there is no data available within Intertek PLC for comparison with the results obtained from this study.

2 Background material

2.1 Support vector machines

Support Vector Machines (SVM) as proposed by Vapnik (Cortes and Vapnik, 1995) were originally intended to be used as a two-class linear classification tool, but it has now been expanded to be applied in non-linearly separable classification with the use of kernel trick. The kernel trick is used to map non-linearly separable data to a higher dimension to enable a class separation (Zhang, 2001), (Li et al., 2023) as is the case with this project. Support Vector Machines (SVM) is a supervised machine learning algorithm which classifies groups in a dataset with the use of a decision hyperplane. It is termed "supervised" because it works with known/labelled classes in a dataset. Data objects are plotted as points on a feature space. Each of the features in the data is depicted by a coordinate. To classify the data, the plotted coordinates are divided with a hyperplane to separate the different classes. 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). This study will investigate the application of SVM as a classification algorithm.

2.1.1 SVM hyperplane

SVM aims at maximizing the margin between the support vectors and the separating hyperplane to aid a binary classification problem (Figure 1). This is why it is termed an optimal separating hyperplane (Gunn, 1998). Support vectors are the data points closest to the separating hyperplane.

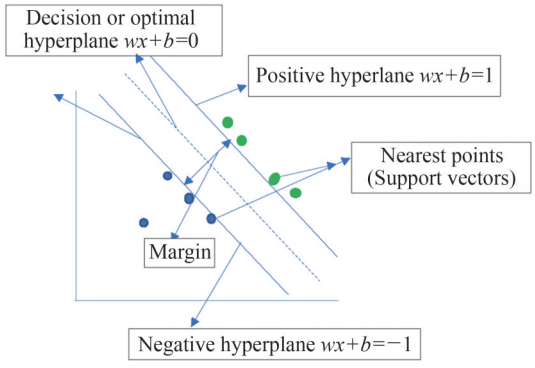

Figure 1 Support vectors and optimal hyperplane

Mathematically, the hyperplane for a linear SVM is described as equation 1.

$$\boldsymbol{w}x + b = 0 \tag{1}$$

where $\boldsymbol{w}$ is the normal vector, x represents the marine fuel samples, b is the intercept and $\frac{b}{\|\boldsymbol{w}\|}$ describes the distance which is perpendicular to the hyperplane from the origin. From Figure 1, it may be observed that, if the positive hyperplane is expressed as $\boldsymbol{w}x + b = 1$ and the negative hyperplane is expressed as $\boldsymbol{w}x + b = -1$, the optimal hyperplane is mathematically shown to maximize the sum of the distance between the closest data point to it on the positive hyperplane and also the distance between it and the closest data point on the negative hyperplane (Liu, 2020) (Wang et al., 2023a). The margin is described as 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 the negative hyperplanes.

2.1.2 SVM margin maximization

To illustrate margin maximization, consider a positive hyperplane $\boldsymbol{w}x^{(p)} + b = 1$ and a negative hyperplane $\boldsymbol{w}x^{(n)} + b = -1$ where $x^{(p)}$ and $x^{(n)}$ are data points in the positive and negative hyperplanes respectively.

The distance between $x^{(p)}$ and the decision hyperplane and $x^{(n)}$ and the decision hyperplane can be determined using equations 2 and 3 respectively (Liu, 2020).

$$\frac{\boldsymbol{w}x^{(p)} + b}{\|\boldsymbol{w}\|} = \frac{1}{\|\boldsymbol{w}\|} \tag{2}$$

$$\frac{\boldsymbol{w}x^{(n)} + b}{\|\boldsymbol{w}\|} = \frac{1}{\|\boldsymbol{w}\|} \tag{3}$$

The margin becomes $\frac{2}{\|\boldsymbol{w}\|}$.

To maximize the margin, $\|\boldsymbol{w}\|$ (the Euclidean norm) needs to be minimized. The quadratic programming process is used to minimize $\|\boldsymbol{w}\|$ given the condition that data points do not fall between the positive and negative points in compliance with the fact that the support vectors are strictly positioned near the hyperplane (Liu, 2020).

In summary, SVM is optimized to locate the support vectors which are responsible for the separation of the margin. The optimization process involves the objective function minimisation and regularization by a constraint and error designation with the Lagrange multiplier. For more details on the Lagrange Multiplier equations see (Bertsekas, 2014) and (Meyer et al., 2003), and the margin maximization process can be seen in (Awad and Khanna, 2015).

2.1.3 Classification with SVM

For a sample of marine fuel data with a vector of wavenumbers x' introduced to the SVM algorithm to be classified and subsequently predicted, we consider equation 4 (Liu, 2020). where $\|\boldsymbol{w}x' + b\|$ is depicted as the distance from x' to the optimal/decision hyperplane.

$$y' = \begin{cases} 1, \text{ if } \boldsymbol{w}x' + b > 0 \\ -1, \text{ if } \boldsymbol{w}x' + b < 0 \end{cases} \tag{4}$$

y equals 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 seen in Figure 1. In the case of the marine fuel sample prediction, y equals 1 (the positive case-HSFO) when the Sulphur content of the fuel is higher than 0.5% by weight and −1 (the negative case-VLSFO) when it is less than 0.5% 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 bigger prediction certainty as the sample is further away from the decision boundary.

2.2 Agglomerative hierarchical clustering algorithm

The agglomerative hierarchical clustering (AHC) algorithm considers each sample of marine fuel as a cluster and constantly merges them at each step of the process forming a hierarchy of clusters (Murtagh and Contreras, 2012). It takes a "bottom-up" approach as it continues to merge smaller clusters into a bigger cluster until all the samples belonging to either HSFO or VLSFO end up in a single cluster or until a set criterion (in this case when 2 clusters are formed) is reached and the process ends (Ahmad and Dang, 2015). The cluster merging process is based on the similarity or distance metric of choice (in this study it is the Euclidean distance). This process builds a dendrogram which is a 2-dimensional illustration of the 'distance between clusters' against the clusters. It shows all the merged clusters at each stage of the process. The merging of the clusters is represented by the branches within the dendrogram (NCSS, 2021). With the aid of the visualized clustering (dendrogram), it is easier to decide on the appropriate number of clusters (marine fuel classes) by making an imaginary 'cut' from the branches of the dendrogram. This is done by considering the magnitude of the change in the merged partition levels in the dendrogram. It is believed that a big difference in the levels or distance between the merged clusters determines the "cut" position to reveal the number of clusters (Everitt and Dunn, 2001). AHC is an unsupervised learning technique (Sreedhar Kumar et al., 2019) because there is no requirement to have priori knowledge of the data distribution. Also, the pre-selection of the number of clusters is not to be defined by the user, which is one of the many advantages of AHC but is also known to suffer from high computational complexity due to cluster merging steps, therefore the process can be too slow with large data sets (Patel and Thakral, 2016). This is not the case with this study due to the small-sized dataset in use however, Principal Component Analysis will be used to reduce the dimensions (186 wave number variables) of the data for the AHC algorithm to reduce the computational complexity while utilising the features that explain the most variance in the data. A similar application of the AHC algorithm in conjunction with principal component analysis was seen in (Zhang et al., 2022) where PCA's ability in pattern recognition aided the clustering of wells with similar steam flooding field applications for enhanced oil recovery process. The wells were assigned to clusters to facilitate effective referencing to operational and performance indicators for future decision-making processes (Zhang et al., 2022). The Ward linkage technique which is based on the sum of the squared distances inside all the clusters (Maklin, 2018) will be used in this study.

2.3 Model performance metrics

The use of accuracy as the performance metric does not present a true picture of the classifier's accuracy when dealing with unbalanced data. A high prediction accuracy of 95% for an unbalanced dataset provides no information about the misclassified minority class. It only provides information to show that most samples have been assigned to the majority class with 95% accuracy (He and Garcia, 2009).

Most classification algorithms assume that all data sets 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 data sets are unbalanced. An unbalanced data is one with a skewed distribution with the number of samples in one class greatly surpassing the other. The classification of such data has been a major hin-

drance to many researchers as most classifiers favour the majority class. (He and Garcia, 2009).

2.3.1 Confusion matrix

The validity of classification models is vital for the subsequent prediction of unknown samples when introduced into supervised algorithms including SVM. Samples can be assigned to the appropriate groups or classes but there is a possibility of misclassification. The prediction results of the model in terms of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) are all captured in a confusion matrix. True Positive Rate (TPR) or Sensitivity, False Positive Rate (FPR), True Negative Rate (TNR) or Specificity, and False Negative Rate (FNR) respectively are all measured based on the confusion matrix values (Westerhuis et al., 2008). See Equations (5)–(7) (Sun, 2009). TPR and TNR portray the correct prediction of positives and negatives respectively while the FPR and FNR stand for the falsely predicted result or prediction errors.

Knowing that the accuracy of a classifier is not an effective measurement criterion for the classification of unbalanced data, it is therefore important to combine a variety of performance metrics to assess the performance of a machine learning model (He and Garcia, 2009) as is the case with this study. A combination of AUROC, Sensitivity and Specificity of the predictions will be used to evaluate the models.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{5}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{(\mathrm{FP} + \mathrm{TN})} \tag{6}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{(\mathrm{FP} + \mathrm{TN})} \tag{7}$$

TPR or sensitivity indicates how many of the positive samples were correctly assigned (Bekkar et al., 2013). It shows that samples are correctly ascribed to a class, while TNR (Specificity) portrays the probability that samples are rightly identified as not being a member of a class as per the chosen threshold (Broadhurst and Kell, 2006), (Sun, 2009).

2.3.2 Receiver operator's characteristic (ROC) curve

A graphical plot of FPR (1-Specificity) against TPR (Sensitivity) for the model generates a curve referred to as the Receiver Operator Characteristics or ROC Curve. The ROC curve, whose origin can be traced back to the use of radar for communication signal detection is an extensively used means of describing a variable's effectiveness in a two-fold classifier as the threshold changes. It is employed as a quality measure for classification models. The area under the ROC curve (AUROC) is required to be greater than the 0.7 (70%) mark for an acceptable classification and above 0.9 for a very good classification (Lantz, 2019). As a non-parametric measure, the ROC curve is considered effective as it incorporates the accuracy measures of sensitivity and specificity of the results regardless of any rudimentary distributions within the population. A low AUROC for instance 0.5 depicts a uniform split of the variable between the two groups of a binary classifier indicating poor discrimination or a random classification while an AUROC of 1 is evidence of a complete separation between the two groups. It is a clear indication that the values of such variables found between the ranges are useful for distinguishing between the classes (Broadhurst and Kell, 2006).

3 Methodology

The industrial spectral data for marine fuel oil was supplied by Intertek PLC. The samples were each scanned about 5 times with the aid of ABB MB3000 Series FTIR Spectrometer generating more than 1 600 NIR observations. The data was positively skewed (Figure 2). The Sulphur content for the majority of the samples within the dataset was below the mean (1.12) which is greater than the median (0.49). It is also unbalanced in favour of samples with low Sulphur content (<0.5%) as more than 80% of samples contain 0-0.5% Sulphur. On the other hand, there were relatively few samples with Sulphur values measuring more than 0.5%. The skewness of the data (1.07) confirms a highly positively skewed (Nagla, 2014) distribution of the marine fuel Sulphur data making it an inherently difficult one to classify (Lantz, 2023) as most machine learning algorithms assume a normal distribution of the data (Awad and Khanna, 2015). A kurtosis of −0.54 also confirms the "flatness" of the data distribution curve (Nagla, 2014). There are 136 HSFO samples and >190 VLSFO samples in the dataset making it more prone to misclassifications of the minority (Awad and Khanna, 2015).

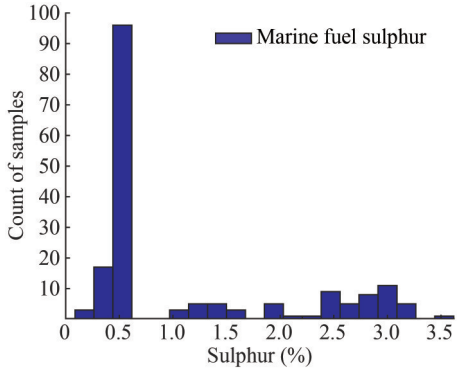

Figure 2 Marine fuel sulphur data's skewed distribution

To predict marine fuel samples based on the spectral information provided, the following steps in Figure 3 were followed for SVM (Workflow for SVM). A similar workflow (Data Cleaning/Removal of invalid signals, Baseline Correction at 4 780 cm^{-1}, MATLAB AHC model Building and Performance Evaluation) was followed for the AHC algorithm. The cluster merging process is shown in the Process Flowchart in Figure 4. These steps are further described in the next sections.

3.1 Data preprocessing

MATLAB Bioinformatics Toolbox was used to read and cleanse the data of redundant and missing values. The average of each sample's five signals was taken to ensure each sample was fully accounted for in the analysis and model building. This reduced the number of samples to >300. 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 of the marine fuel samples. 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). In this study, these were mitigated using baseline shift correction at 4 780 cm^{-1} (the wavenumber where these effects were seen) and normalising data with the area under the curve set to one and followed by standardisation to ensure all samples' signals have equal weighting in the model building. The resultant effect on the model with both normalisation methods will be compared.

The marine fuel data was normalised (Figure 5) with the area under the curve set to one after carrying out a baseline shift to reduce the influence of the signal shifts on the normalised data restoring all the signal peaks to a common baseline (Fanali et al., 2017) is hereafter referred to as Data A. The normalisation process of setting the area under the curve to one scales the data to have the integral of the wavenumber values to be equal to one. This data was further standardised by subtracting the mean from each sample vector before scaling it by its standard deviation. It is an ideal practice to normalise data by setting the area under the curve to one in such a way that all the signals are modified and summed to one thereby creating more uniform data with equal contribution from each sample without any bias from data points with larger wavenumber values 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 data was also standardised (Figure 6) without setting the area under the curve to one. This is hereafter referred to as Data B and will be used to build a second model for comparison purposes. Although tighter spectra were seen with the standardised data, it could be seen that the difference between the two normalisation methods was insignificant.

Normalising by setting the area under the curve to one was completed by using the trapezoidal integration function within the MATLAB environment to generate the area whilst setting the area under the curve to one by dividing all the signal values by the area output.

All exploratory data analysis and cleaning processes including the removal of spiked/abnormal signals were focused on the 4 000–4 800 cm^{-1} wavenumber range which is associated with crude and marine fuel's signal peaks/absorption in the combination region range. This is also in line with Intertek PLC's operating range. The peaks around 4 200 cm^{-1} and 4 300 cm^{-1} and between 4 300 cm^{-1} and 4 400 cm^{-1} are the most significant peaks used in characterising the sulphur content of marine fuel oil (Lammoglia and de Souza Filho, 2011).

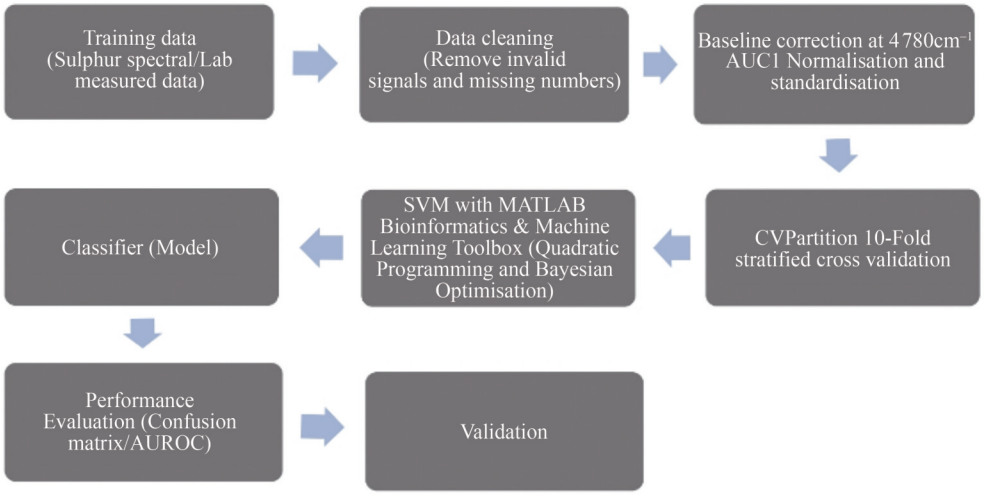

Figure 3 Workflow for SVM

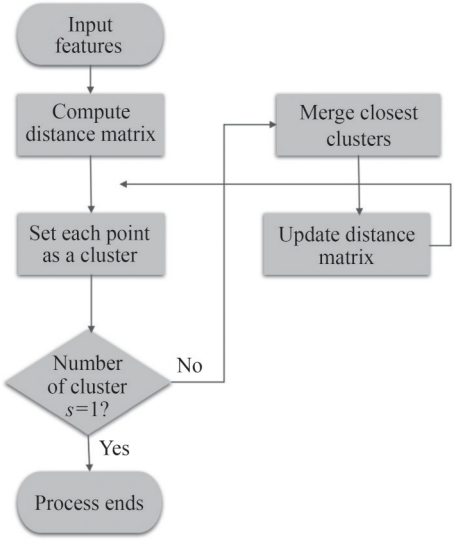

Figure 4 Process flowchart for AHC

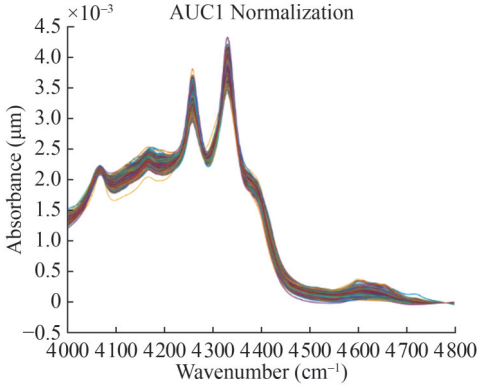

Figure 5 Normalised by setting area under the curve to 1

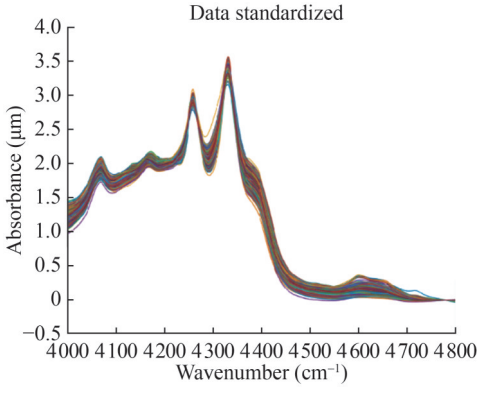

Figure 6 Marine fuel data standardised

3.2 Model building and cross-validation

The Statistics & Machine Learning and Phased Array System Toolbox from Mathworks' MATLAB were used to build models with the Support Vector Machines algorithm using the Radial Basis Function (RBF) kernel for binary classification. The Bayesian Optimisation (Gelbart et al., 2014) within the toolbox was used to find the hyperparameters including the kernel and parameter C that minimise the objective function (cross-validation loss). This model was trained with soft margin minimization through SVM's quadratic programming to aid the objective function minimization and maximize the margin separation between the support vectors and the hyperplane (Mathworks, 2022). There were 30 function evaluations for Data A and B respectively for the selection of minimum objective functions with the aid of Bayesian optimisation. This optimizes the C parameter (depicted here by the Box Constraint to determine and maximize the margin separation between the support vectors and the hyperplane thereby setting the appropriate cost for misclassification of high Sulphur and low sulphur samples in the marine fuel data set (Mathworks, 2022). It revealed the estimated objective function of 0.198 54 working with a classification error cost function illustrated with a Box Constraint of 26.153 and a KernelScale of 49.629 for Date A (Figure 7). Data B showed an estimated objective function value that was optimised to 0.220 83 with the best estimated feasible point (according to models) showing the Box Constraint 0.953 63 and a KernelScale of 1.591 8 (Figure 8). A total of 122 Support vectors were used to affect the separation of the classes (HSFO and VLSFO).

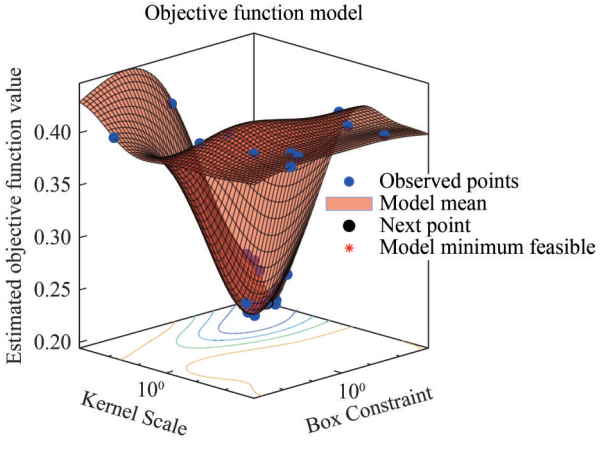

Figure 7 Objective function minimisation by bayesian optimisation Data A

Given the limited size and unbalanced nature of the data, and to prevent overfitting the model (Wang et al., 2023b), it was partitioned and cross-validated with MATLAB's KFold group stratified partitioning technique to ensure a random distribution of both HSFO and VLSFO fuel classes while assigning approximately the same number of samples and class proportions in each fold ensuring that samples do not appear in more than one fold. This method is specifically known to be applied to mitigate the data imbalance effect in modelling. The data was split into 10 folds to allow the utilisation of 9 infold samples in the training set leaving

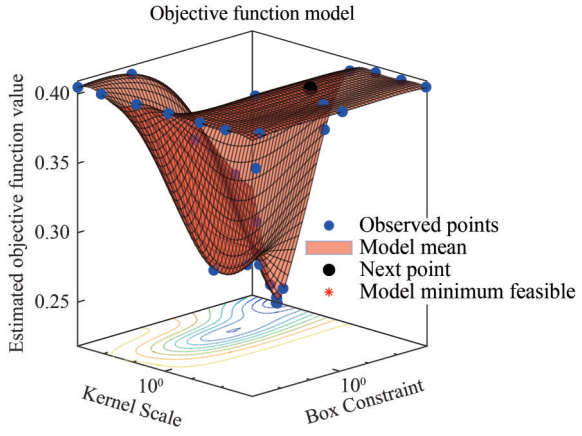

Figure 8 Objective function minimisation by bayesian optimisation Data B

1-fold for validation at each iteration. This will form the internal validation process. The internal validation of the SVM models was carried out via MATLAB's KFoldPrediction method used to predict the out-of-fold samples with the model built without the samples to be predicted. To be precise, "KfoldPredict" calculates the predictions for the samples in the first out-of-fold group (group a) of samples with the aid of the first model built without group a, then goes on to calculate the predictions for the samples in the second out of fold group (group b) with the aid of the second model built without group b and so on. In a nutshell, the software evaluates a response for every sample in the marine fuel dataset using the model trained without that particular sample. To further mitigate the effect of data imbalance on the model, the inverse hyperbolic sine of the spectral data was used to to create a more symmetric and normal data distribution (Liping et al., 2017).

80% of the samples were used to train the classifier while the remaining 20% of the samples were set aside for model validation. The performance of the models was assessed based on the confusion matrix (Broadhurst and Kell, 2006) values: True Positive Rate (Sensitivity), True Negative Rate (Specificity), and Area under the Receiver Operating Characteristics curve (AUROC) value (Spackman, 1989).

To validate and compare the results, MATLAB Statistics & Machine Learning and Phased Array System Toolbox were used to build models with the AHC clustering algorithm. Principal Component Analysis (PCA) was used to reduce the dimensionality of the marine fuel data set. To achieve this, PCA was used to find the group of orthogonal sample vectors that capture most of the data variance in the data thereby reducing 186 wavenumber-data to principal components, keeping all the crucial information but with lower dimension data (2 Principal Components) and maintaining a high percentage (90%) (PC1 & 2) of explained variance in the data. Figure 9 shows a plot of percentage cumulative variance captured as a function of the principal component number. It could be seen that >90% of the important information depicted by the 20 useful components was captured in the first 3 PCs with the PC1 and 2 containing a cumulative variance of >85%. To investigate the effects of the choice of the principal components to be included in the model, a trial of different combinations of the principal components was carried out. Each combination is used to build the AHC models to verify the combination with a more accurate true prediction. SVM models were built without a dimensionality reduction process.

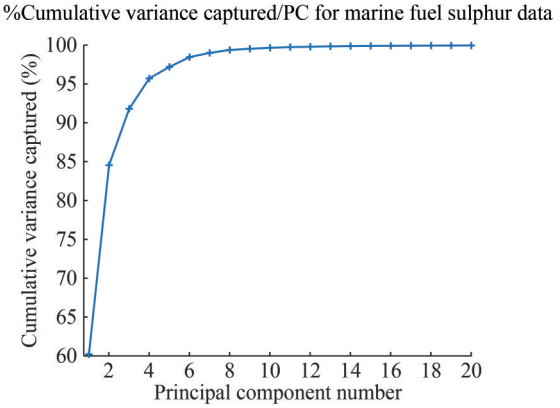

Figure 9 %Cumulative variance captured vs number of principal components

In addition to the first two PCs selected as the clustering modelling features (PC1 & 2 with 90% of variations in the dataset explained), other PCs (PC1 & 3, PC2 & 3) were also used to build 2 other models with the AHC algorithm to compare the models' accuracy and performances with the marine fuel's spectral information correlated with laboratory-measured Sulphur data.

The data was normalised with the area under the curve set to 1 before standardising. The termination criterion for the AHC algorithm was set to a maximum of 2 clusters based on the visual assessment of the data.

4 Results and discussion

This section details the results obtained from SVM and AHC algorithms. Model validations were carried out using the unseen (20%) data samples in addition to the rest of the samples as specifically requested by Intertek PLC due to data scarcity.

4.1 SVM models

The use of Bayesian optimisation created an optimum hyperplane to separate high and low Sulphur fuel oil while minimising the classification error. This successfully yielded a result with a True Positive Rate (Sensitivity) for Data A (78.1% HSFO) and Data B (72.5% HSFO) for the internally

(cross) validated data. A similar result (Figures 10 and 11) for the externally validated test set (with the entire data sets) revealed a true positive rate of 72.3% for Data A indicating that 72.3% of HSFO samples were accurately predicted and 71.3% were correctly predicted from Data B respectively. This shows that almost an average of 75% of the marine fuel oil samples at >0.5% Sulphur content was predicted accurately from Data A while an average of 72% high Sulphur fuel oil samples (standardised) were correctly predicted with a Specificity (TNR) of 82.7% VLSFO for Data A and 81.3% VLSFO for Data B were recorded. This indicates that 81% of samples with <0.5% of Sulphur content was correctly predicted with the data normalised with the area under the curve set to 1 before standardising while the result was slightly less with the standardised data. The external validation revealed a TNR of 82.7 showing that 82.7% of VLSFO fuel oils were accurately predicted with Data A while 79.8% of VLSFO samples were accurately predicted with Data B. It is interesting to note that the inverse hyperbolic sine of the data generated a prediction at a similar true positive rate (73.9%) with the original data. The true negative rate (78.4%) and a FNR of 21.6% show that HSFO fuel samples were falsely identified as VLSFO. A false positive rate of 26.1% indicates that the inverse hyperbolic sine of the marine fuel data did not improve the prediction results.

The cross-validation accuracy of 81% and 19% training error were recorded for Data A while a cross-validation training accuracy of 78% with a training error rate of 22% was recorded for Data B showing that normalising the data with the area under the curve set to a unit had a slightly better influence on the ability of SVM to classify marine fuel sulphur data. The difference, however, is insignificant.

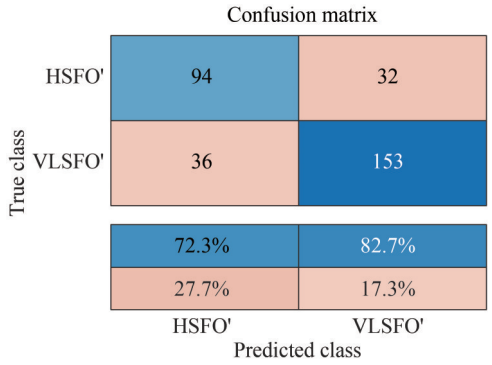

Figure 10 External validation (Data A)

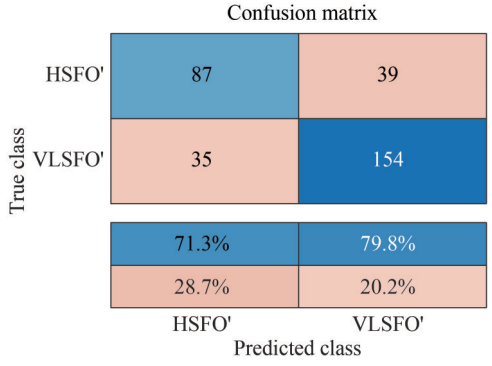

Figure 11 External validation (Data B)

A false-positive rate (FPR) of 21.9% indicates that 21.9% of marine fuel oil with less than 0.5% Sulphur content (VLSFO) was incorrectly assigned to HSFO class while a 17.3% false-negative rate (FNR) shows that 17.3% of HSFOs were falsely grouped as VLSFOs for Data A. Similarly, an FPR of 27.5% shows that 27.5% of VLSFOs were incorrectly grouped as HSFOs while an FNR of 18.7% of HSFOs indicates that 18.7% of HSFO samples were inaccurately assigned to the VLSFO class in Data B.

The external validation with a 27.7 FPR indicates that 27.7% of VLSFO samples were falsely predicted as HSFO samples while a 17.3% FNR shows that 17.3% of HSFOs were falsely predicted as VLSFOs with Data A.

With Data B, a 28.7 FPR indicates that 28.7% of VLSFO samples were falsely predicted as HSFO samples while a 20.2 FNR shows that 20.2% of HSFOs were falsely predicted as VLSFOs. Although an acceptable result, the false predictions showcase the reliance of the machine learning model's accuracy on the data size. Several factors including small data size affect the accuracy due to the limitations in the data availability for high Sulphur fuel oil. The provision of more data would enable the improvement of the model (Mehta and Kundra, 2024)

The Area under the ROC curve (AUROC) of 0.86 for internally validated Data A and 0.83 for Data B confirms the accuracy of the classification of marine fuel based on the NIR spectral fingerprints correlated with the laboratory-measured Sulphur composition by weight of the marine fuel oil. The result is not far from the Area under the ROC of 0.83 and 0.82 with the externally validated test set (with the entire data set including the 20% unseen samples). This indicates that the true positive rates at different thresholds were at an average of >0.80 showing the potential of a good prediction of high Sulphur fuel oil samples with the Support Vector Machines model. It also shows that the models have a >80% chance that the models will accurately rank one randomly chosen HSFO sample (which is a positive case here) over a randomly selected VLSFO sample (a negative class).

4.2 AHC models

The dendrogram (Figure 12) of marine fuel data (PC1 & 2) normalised with the area under the curve set to oneshowed the 2 clusters discovered by the algorithm. The merging of clusters is depicted by the branching of the lines. An imaginary cut from the widest part of the dendrogram will clearly show the 2 clusters. These clusters depict the classes belonging to high and low Sulphur fuel oil respectively.

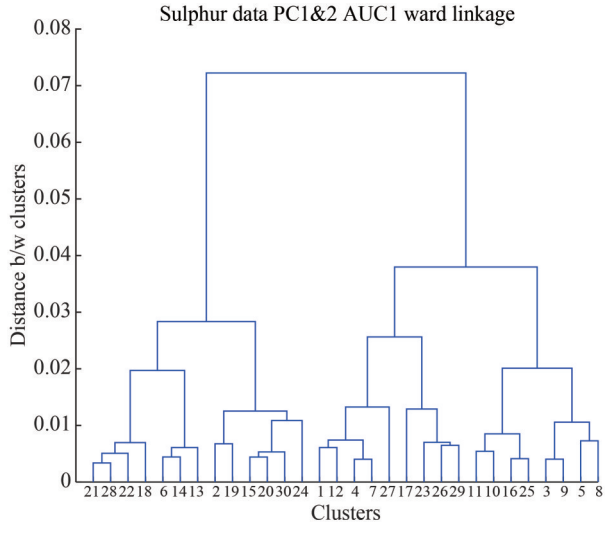

Figure 12 AHC dendrogram for marine fuel showing 2 clusters discovered

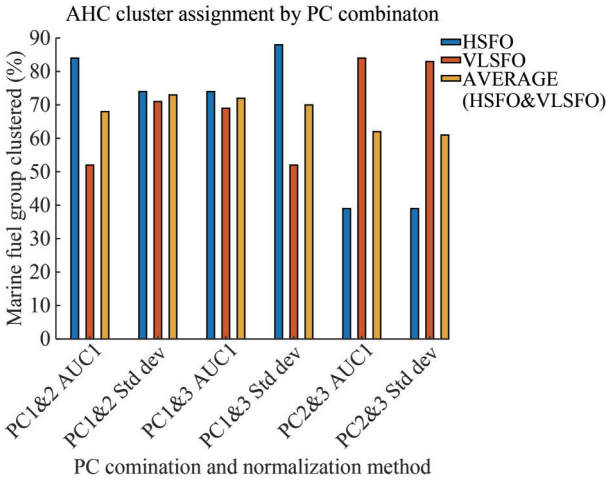

Figure 13 AHC cluster results

Figure 13 shows that a more effective clustering result was achieved with AHC when compared with the SVM algorithm. A higher percentage of HSFO (up to 88% PC1 & 2 AUC1) and VLSFO (up to 84% PC2 & 3 AUC1) were accurately clustered. It could be seen that PC1 & 2 for data scaled by standard deviation presents a consistent result for both HSFO and VLSFO with an average of 73% for PC1 & 2 and followed closely by PC1 & 3 for data normalized with the area under the curve set to a unit at 72%. PC2 & 3 showed a poor clustering of the HSFO fuel group at 39% for PC2 & 3 AUC1 and standard deviation normalisation methods.

In summary, looking at the average clustering outcome for both marine fuel groups, the result indicates a slightly higher accuracy with the model built with PC1 & 2 using the data scaled by standard deviation and PC1 & 3 using the data normalized with the area under the curve set to a unit with AHC algorithm. The difference between AUC1 and standardised data is insignificant. Based on this, it could be deduced that the normalisation method did not have a huge influence on the AHC model results. It is also worth mentioning that the unsupervised AHC algorithm outperformed the SVM without the need for prior knowledge of the data structure or classes without major parameter tuning or optimisation. The summary of the results is seen in Table 1.

Similarly, HSFO samples were more accurately predicted at 88% with the AHC algorithm with the standardised data (PC1 & 3) while the VLSFO samples were more accurately predicted at 84% (PC2 & 3) with the data that has the area under the curve set to one before normalising. This indicates that the AHC algorithm was a more effective technique for the prediction of HSFO and VLSFO samples with Intertek PLC's industrial dataset.

Table 1 summarises TPR and TNR for AHC and SVM models and highlights the best predictions for HSFO and VLSFO

Fuel type	SVM (TPR)	AHC (TPR)	SVM (TNR)	AHC (TNR)	AUC1	Standardised	PC used	Comment
HSFO	72.3				x		N/A	Best prediction SVM (HSFO)
HSFO	71.3					x	N/A	
VLSFO			82		x		N/A	Best prediction SVM (VLSFO)
VLSFO			79.8			x	N/A	
HSFO		74				x	PC1 & 2	
HSFO		88				x	PC1 & 3	Best prediction AHC (HSFO)
HSFO		84			x		PC1 & 2	
HSFO		74			x		PC1 & 3	
VLSFO		71	71			x	PC1 & 2	
VLSFO		83	83			x	PC2 & 3	
VLSFO		84			x		PC2 & 3	Best prediction AHC (VLSFO)

5 Conclusion

AHC, an unsupervised learning algorithm outperformed SVM, a supervised learning algorithm by successfully clustering 2 classes of marine fuel (HSFO and VLSFO) which is in line with the data supplied by Intertek PLC. Although there were class labels supplied with the data, the unsupervised AHC algorithm was not adversely affected by this information which was not required. The termination criterion was set to end when 2 clusters were formed. The models showed higher predictive ability than the SVM models for HSFO with PC1 & 3 (Standardised) and PC1 & 2 (with the data normalised with the area under the curve set to one). For VLSFO models, SVM showed an acceptable result with the data normalised with the area under the curve set to one before standardising and also with the standardised data. AHC models also showed higher predictive ability than the SVM models. Although insignificant, the data with the area under the curve set to a unit had a slightly better influence on the ability of SVM to classify marine fuel Sulphur data, therefore both normalisation methods effectively influenced the prediction of marine fuel oils. With the AHC algorithms, the choice of the relevant principal components played a role in the effective clustering and prediction of the marine fuel oil samples.

The classification, clustering and prediction of marine fuel oil based on their Sulphur characteristics and spectral information are vital to marine vessel operators. This is a positive and promising result considering that this is an initial attempt by Intertek PLC to utilise a classification and/or clustering algorithm for the prediction of marine fuel oils with high Sulphur content to provide a prompt and informed decision-making resource for their customers concerning the fuels to be used onboard ocean-going vessels, especially in ECAs. It is an interesting result considering that NIR elemental measurements (e.g. Sulphur) are inferred predictions as the infrared spectral data is specifically linked to functional groups attached to Sulphur. The predictions are aimed at enabling the vessel operators to remain compliant with IMO regulations when choosing fuel oils instead of relying on time-consuming laboratory methods. However, there is room for more investigations into a direct prediction of the Sulphur contents of the fuel oil samples.

Nomenclature

NIR	Near infrared
ASTM	American society for testing & materials
ISO	International standard organisation
RBF	Radial basis function
TP	True positives
TN	True negatives
FP	False positives
TPR	True positive rate
TNR	True negative rate
FPR	False positive rate
FNR	False negative rate
ROC	Receiver operating characteristics
AUROC	Area under the receiver operating characteristics curve

Funding This work was supported by Newcastle University and the Engineering and Physical Sciences Research Council (EPSRC) [grant numbers 2020/21 DTP: ref.EP/T517914/1].

Competing interests The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Ahmad H, Dang S (2015) Performance Evaluation of Clustering Algorithm Using Different Dataset. International Journal of Advance Research in Computer Science and Management Studies, 8

Al Ibrahim E, Farooq A (2021) Prediction of the Derived Cetane Number and Carbon/Hydrogen Ratio from Infrared Spectroscopic Data. Energy & Fuels 35(9): 8141-8152. https://doi.org/10.1021/acs.energyfuels.0c03899

Awad M, Khanna R (2015) Support Vector Machines for Classification. Efficient Learning Machines, 39-66

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, 48-49

Bekkar M, Djemaa HK, Alitouche TA (2013) Evaluation Measures for Models Assessment over Imbalanced Data Sets. J Inf Eng Appl 3(10)

Bertsekas DP (2014) Constrained Optimization and Lagrange Multiplier Methods. Academic press. https://doi.org/10.1016/C2013-0-10366-2

Bilgili L (2021) Life Cycle Comparison of Marine Fuels for Imo 2020 Sulphur Cap. Science of The Total Environment 774: 145719. https://doi.org/10.1016/j.scitotenv.2021.145719

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

Broadhurst DI, Kell DB (2006) Statistical Strategies for Avoiding False Discoveries in Metabolomics and Related Experiments. Metabolomics 2(4): 171-196

Christopher J, Patel MB, Ahmed S, Basu B (2001) Determination of Sulphur in Trace Levels in Petroleum Products by Wavelength-Dispersive *X*-Ray Fluorescence Spectroscopy. Fuel 80(13): 1975-1979. https://doi.org/10.1016/S0016-2361(00)00213-1

Ciaburro G, Joshi P (2019) 1.6 Normalization. Python Machine Learning Cookbook (2nd Edition)

Concawe (2016) Marine Fuel Facts, 2022 (10 November)

Corbett JJ, Winebrake JJ, Green EH, Kasibhatla P, Eyring V, Lauer A (2007) Mortality from Ship Emissions: A Global Assessment. Environmental Science & Technology 41(24): 8512-8518. https://doi.org/10.1021/es071686z

Cortes C, Vapnik V (1995) Support-Vector Networks. Machine

learning 20(3): 273-297. http://dx.doi.org/10.1007/BF00994018
Cullinane K, Bergqvist R (2014) Emission Control Areas and Their Impact on Maritime Transport. Transportation Research Part D: Transport and Environment 28: 1-5. https://doi.org/10.1016/j.trd.2013.12.004
Dadi HS, Pillutla GM (2016) Improved Face Recognition Rate Using Hog Features and Svm Classifier. IOSR Journal of Electronics and Communication Engineering 11(04): 34-44. http://dx.doi.org/10.9790/2834-1104013444
Deng F, Guo S, Zhou R, Chen J (2015) Sensor Multifault Diagnosis with Improved Support Vector Machines. IEEE transactions on automation science and engineering 14(2): 1053-1063. https://doi.org/10.1109/TASE.2015.2487523
Everitt BS, Dunn G (2001) 6.2 Agglomerative Hierarchical Clustering Techniques. Applied Multivariate Data Analysis (2nd Edition)
Eyring V, Isaksen ISA, Berntsen T, Collins WJ, Corbett JJ, Endresen O, Grainger RG, Moldanova J, Schlager H, Stevenson DS (2010) Transport Impacts on Atmosphere and Climate: Shipping. Atmospheric Environment 44(37): 4735-4771. https://doi.org/10.1016/j.atmosenv.2009.04.059
Fan L, Shen H, Yin J (2023) Mixed Compliance Option Decisions for Container Ships under Global Sulphur Emission Restrictions. Transportation Research Part D: Transport and Environment 115: 103582. https://doi.org/10.1016/j.trd.2022.103582
Fanali S, Haddad PR, Poole CF, Riekkola M-L (2017) 21.3.3 Normalization. Liquid Chromatography-Fundamentals and Instrumentation, Volume 1 (2nd Edition)
Gelbart MA, Snoek J, Adams RP (2014) Bayesian Optimization with Unknown Constraints. arXiv preprint arXiv: 1403.5607. https://doi.org/10.48550/arXiv.1403.5607
Gu Y, Wang Y, Iris Ç (2025) Integrated Green Technology Adoption, Ship Speed Optimization and Slot Management for Shipping Alliance under Emission Limits and Uncertain Fuel Prices. Journal of Cleaner Production 494: 144939. https://doi.org/10.1016/j.jclepro.2025.144939
Gunn SR (1998) Support Vector Machines for Classification and Regression. ISIS technical report 14(1): 5-16
Hassellöv IM, Turner DR, Lauer A, Corbett JJ (2013) Shipping Contributes to Ocean Acidification. Geophysical Research Letters 40(11): 2731-2736. https://doi.org/10.1002/grl.50521
He H, Garcia EA (2009) Learning from Imbalanced Data. IEEE Transactions on knowledge and data engineering 21(9): 1263-1284. https://doi.org/10.1109/TKDE.2008.239
Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B (1998) Support Vector Machines. IEEE Intelligent Systems and their applications 13(4): 18-28. https://doi.org/10.1109/5254.708428
Huang J, Romero-Torres S, Moshgbar M (2010) Practical Considerations in Data Pre-Treatment for Nir and Raman Spectroscopy, American Pharmaceutical Review. Dostopno na: http://www. american pharmaceuticalreview. com/Featured-Articles/116330-Practical-Consi derations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy/.[Dostop: 10-Sep-2019]
IHMMarineSurveys (2020) Fuel Oil Sulphur Testing and Analysis, 2023 (20 July)
Ju H-j, Jeon S-k (2022) Effect of Ultrasound Irradiation on the Properties and Sulfur Contents of Blended Very Low-Sulfur Fuel Oil (Vlsfo). Journal of Marine Science and Engineering 10(7): 980. https://doi.org/10.3390/jmse10070980
Kapoutsis E, Theodoulidis B, Saraee M (2024) Svm Categorizer: A Generic Categorization Tool Using Support Vector Machines. Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, 1109-1112.
Kuzu SL, Bilgili L, Kiliç A (2021) Estimation and Dispersion Analysis of Shipping Emissions in Bandirma Port, Turkey. Environment, Development and Sustainability 23(7): 10288-10308. https://doi.org/10.1007/s10668-020-01057-6
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
Lantz B (2019) 10.1.5 Visualizing Performance Tradeoffs with Roc Curves. Machine Learning with R (3rd Edition), 331-332
Lantz B (2023) Machine Learning with R (4th Edition) -Learn Techniques for Building and Improving Machine Learning Models, from Data Preparation to Model Tuning, Evaluation, and Working with Big Data
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
Liping W, Xuelong H, Jiang N (2017) Robust Time Delay Estimation Based on Asinh Transform under α -Stable Noises. 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI) 162-166. https://doi.org/10.1109/ICEMI.2017.8265932
Liu Y (2020) Python Machine Learning by Example (3rd Edition). Packt Publishing
Maklin C (2018) Hierarchical Agglomerative Clustering Algorithm: Example in Python, 2021(21 July)
Mathworks (2022) Fitcsvm, 2022(June 06)
Mehta S, Kundra D (2024) Combining Cnn and Svm for Robust Cattle Disease Classification in Veterinary Applications. 2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT) 1-5. https://doi.org/10.1109/IC-SIT63503.2024.10862162
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
Murtagh F, Contreras P (2012) Algorithms for Hierarchical Clustering: An Overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(1): 86-97. https://doi.org/10.1002/widm.53
Nagla JR (2014) Statistics for Textile Engineers
NCSS (2021) Hierarchical Clustering/Dendrograms, 2021 (09 September)
Patel KA, Thakral P (2016) The Best Clustering Algorithms in Data Mining. 2016 International Conference on Communication and Signal Processing (ICCSP) 2042-2046. https://doi.org/10.1109/ICCSP.2016.7754534
Saldana DA, Starck L, Mougin P, Rousseau B, Ferrando N, Creton B (2012) Prediction of Density and Viscosity of Biofuel Compounds Using Machine Learning Methods. Energy & fuels 26(4): 2416-2426
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
Spackman KA (1989) Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning. Proceedings of the sixth international workshop on Machine learning, 160-163. https://doi.org/10.1016/B978-1-55860-036-2.50047-3
Sreedhar Kumar S, Madheswaran M, Vinutha B, Manjunatha Singh H, Charan K (2019) A Brief Survey of Unsupervised Agglomerative Hierarchical Clustering Schemes. Int J Eng Technol (UAE) 8(1): 29-37
Stratiev D, Dinkov R, Petkov K, Stanulov K (2010) Evaluation of

Crude Oil Quality. Petroleum & Coal 52(1): 35-43
Sun D-W (2009) 4.3 Evaluation of Classification Performances. Infrared Spectroscopy for Food Quality Analysis and Control
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
Van TC, Ramirez J, Rainey T, Ristovski Z, Brown RJ (2019) Global Impacts of Recent Imo Regulations on Marine Fuel Oil Refining Processes and Ship Emissions. Transportation Research Part D: Transport and Environment 70: 123-134. https://doi.org/10.1016/j.trd.2019.04.001
Wang H, Hu L, Zhang Y (2023a) Svm Based Imbalanced Correction Method for Power Systems Transient Stability Evaluation. ISA Transactions 136: 245-253. https://doi.org/10.1016/j. isatra. 2022. 10.039
Wang Q, Chen D, Li M, Li S, Wang F, Yang Z, Zhang W, Chen S, Yao D (2023b) A Novel Method for Petroleum and Natural Gas Resource Potential Evaluation and Prediction by Support Vector Machines (Svm). Applied Energy 351: 121836. https://doi.org/10.1016/j.apenergy.2023.121836
Westerhuis JA, Hoefsloot HC, Smit S, Vis DJ, Smilde AK, van Velzen EJ, van Duijnhoven JP, van Dorsten FA (2008) Assessment of Plsda Cross Validation. Metabolomics 4(1): 81-89
Workman J (2001) Handbook of Organic Compounds: Nir, Ir, Raman and Uv-Vis Spectra Featuring Polymers and Surfactants (a 3-Volume Set). 3. Ir and Raman Spectra. Academic Press
Zhang N, Wei M, Bai B, Wang X, Hao J, Jia S (2022) Pattern Recognition for Steam Flooding Field Applications Based on Hierarchical Clustering and Principal Component Analysis. ACS Omega 7(22): 18804-18815. http://dx.doi.org/10.1021/acsomega.2c01693
Zhang T (2001) An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Ai Magazine 22(2): 103-103. https://doi.org/10.1017/CBO9780511801389
Zis TP, Cullinane K (2020) The Desulphurisation of Shipping: Past, Present and the Future under a Global Cap. Transportation Research Part D: Transport and Environment 82: 102316. https://doi.org/10.1016/j.trd.2020.102316