Fuel 334 (2023) 126696

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Review article

Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates

Elise Lunde Gjelsvik [a,*], Martin Fossen [b], Kristin Tøndel [a]

[a] Norwegian University of Life Sciences, Faculty of Science and Technology, Ås, Norway
[b] SINTEF AS, Trondheim, Norway

ARTICLE INFO

Keywords:
Gas hydrates
Machine learning
FT-ICR MS
Chemometrics
Crude oil

ABSTRACT

Gas hydrates represent one of the main flow assurance challenges in the oil and gas industry as they can lead to plugging of pipelines and process equipment. In this paper we present a literature study performed to evaluate the current state of the use of machine learning methods within the field of gas hydrates with specific focus on the oil chemistry. A common analysis technique for crude oils is Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) which could be a good approach to achieving a better understanding of the chemical composition of hydrates, and the use of machine learning in the field of FT-ICR MS was therefore also examined. Several machine learning methods were identified as promising, their use in the literature was reviewed and a text analysis study was performed to identify the main topics within the publications. The literature search revealed that the publications on the combination of FT-ICR MS, machine learning and gas hydrates is limited to one. Most of the work on gas hydrates is related to thermodynamics, while FT-ICR MS is mostly used for chemical analysis of oils. However, with the combination of FT-ICR MS and machine learning to evaluate samples related to gas hydrates, it could be possible to improve the understanding of the composition of hydrates and thereby identify hydrate active compounds responsible for the differences between oils forming plugging hydrates and oils forming transportable hydrates.

1. Introduction

Gas hydrates are crystalline structures where smaller guest molecules are trapped in cages formed by water molecules that are held together by hydrogen bonds [1]. Gas hydrates are among the main flow assurance issues when producing oil and gas, especially subsea or in cold locations, because they can lead to complete blockage (plugging) of pipelines and process equipment forcing the operator to shut down the production [2]. The most common, yet very conservative, hydrate strategy states that the positive driving forces for hydrate formation, i.e. high pressure and low temperature, should be avoided. In practice this requires determination of the thermodynamic region where hydrate formation occurs in order to keep the system outside this pressure–temperature region. [3].

For hydrate inhibition on the other hand, the most common strategy is currently the use of thermodynamic inhibitors (THIs). These inhibitors shift the hydrate curve towards higher pressures at hydrate inducing temperatures, enabling production at lower temperatures without the formation of gas hydrates [4,5]. Common inhibitors are organic chemicals, such as methanol and monoethylene glycol (MEG) dosed at concentrations of 20%–50% of the mass relative to the water produced [4]. The premise of their application is that gas hydrate formation is expected, and therefore the inhibitors are always present in the pipelines. Another promising strategy for hydrate management is the injection of low dose hydrate inhibitors (LDHI) [6]. The two main types of LDHIs are the kinetic hydrate inhibitors (KHI) which alter the kinetics during the hydrate formation, and the anti-agglomerants (AAs) which alter wettability of the hydrate particles and prevent them from sticking together. A typical concentration for an LDHI injection is 0.1–1 wt % relative to the water phase [4,7]. For the AAs the purpose is to form a slurry of gas hydrates dispersed in the oil phase that can be transported through the pipelines without the particles aggregating together or depositing to the pipe wall. However, for an AA to be efficient, it must be surface active and able to adsorb to the surface or interact with the hydrate cages of the dispersed hydrate particles. The purpose of KHIs, on the other hand, is to delay the formation of hydrates long enough to reach the storage facility without causing blockage [8]. The KHI binds to the hydrate surface, decreasing the crystal formation process by preventing the growth of hydrate crystals nuclei [9].

* Corresponding author.
E-mail address: elise.lunde.gjelsvik@nmbu.no (E.L. Gjelsvik).

https://doi.org/10.1016/j.fuel.2022.126696
Received 22 July 2022; Received in revised form 12 October 2022; Accepted 7 November 2022
Available online 21 November 2022

However, through laboratory experiments spurred by field experience, it became evident that some crude oils did not experience plugging when gas hydrates were formed [10]. Instead, the hydrates behaved more like dry particles that could be transported without any issues [11]. The explanation set forth was that some crude oils contain naturally occurring components that interact with the gas hydrates rendering the surface of the particles hydrophobic. One hypothesis is that these components have the ability to adsorb to the hydrate surface, preventing agglomeration of hydrates and the potential plugging of the pipeline [12]. Another hypothesis is that parts of a molecule, for example butyl/pentyl groups, penetrate open cavities on the hydrate surface (of $5^{12}\,6^4$ SII cages) and can become embedded in the surface as the hydrate grows around the alkyl groups [4]. The current status of the search for the type and structures of natural hydrate inhibitors is that they have not yet been characterised in detail [2,11–13]. Some previous studies have suggested that these natural inhibitors may be contained in the petroleum acid fraction [11,14–17] which has been shown to include a large amount of naphthenic compounds. Borgund et al. [15] and Erstad et al. [18] showed experimentally the anti-agglomerating properties of some petroleum acid fractions.

Similarly, the asphaltene fractions are known to possess self-agglomerating properties that can stabilise some crude oil systems [19] and some asphaltenes can alter the plugging potential of crude oils [20,21]. It has been shown that the asphaltene fractions able to stabilise systems prone to form transportable slurries are often more polar, with higher oxygen content, higher acidity and lower double bound equivalents (DBEs) [22]. Other studies have suggested that the possible hydrate activity of asphaltenes is related to their sulfoxide content [23].

The overall goal of this review was to establish a baseline for the current status of the use of machine learning in the field of petroleum gas hydrates. A part of this study was to identify work related to naturally occurring hydrate inhibitors in crude oils where machine learning methods have been used. It was, however, shown that this research was extremely limited, resulting in only one publication [24]. Therefore, the methodologies described are related to the thermodynamic aspects of gas hydrates and the chemical analysis of crude oils. Fourier Transform Ion Cyclotron Mass Spectrometry (FT-ICR MS) has a high mass accuracy which could be utilised for analysis of properties related to gas hydrates. FT-ICR MS was therefore included in this review to establish a link between aspects of gas hydrates and analysis of crude oils.

2. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS)

The complex mixtures of crude oils and the relatively high masses of their components make detailed identification difficult with most mass spectrometers. However, with the high mass accuracy of FT-ICR MS, more detailed analysis of crude oil samples are possible [25,26]. In FT-ICR MS the mass-to-charge (m/z) ratio of ions are determined based on the cyclotron frequency of the ions in a fixed magnetic field. The mass accuracy for FT-ICR MS is sub ppm and the mass spectral resolution can be above 10 million (at m/z = 400), which allows identification of a large number of different polar groups [27–29]. In an FT-ICR MS analysis, ions are detected simultaneously within a detecting interval by the ion cyclotron resonance frequency they produce when they rotate in a magnetic field. This provides the increase in signal-to-noise ratio compared to traditional mass spectrometers.

There are several different ionisation techniques to be used in combination with FT-ICR MS. For crude oils, the most common are electrospray ionisation (ESI) and atmospheric pressure photo ionisation (APPI) as they ionise polar compounds efficiently [27,30]. ESI is achieved by applying a high voltage to a liquid passing through a capillary tube inducing highly charged droplets [31]. In positive mode, formic acid is added to the solution aiding ionisation, while in negative mode ammonium hydroxide is added resulting in lower background noise. APPI is performed by exposing the analytes to photons emitted from a UV lamp [27] and in positive mode, both molecular ($[M^{+\cdot}]$) and protonated ions ($[M+H]^+$) are generated. During negative mode, the ions of the molecular species are produced by either proton abstraction or adduct formation. The predominant ions are the molecular species ion ($[M-H]^-$), which is the ion corresponding to the fatty acids ($R_n - COO^-$) present in the sample [31]. APPI is sensitive to aromatic compounds and sulphur containing compounds.

FT-ICR MS has previously been used widely for crude oil characterisation [27,32–38]. For instance, Qian et al. [39,40] showed that positive and negative mode ESI-FT-ICR MS are able to characterise different aspects of crude oils. In negative mode it was identified over 3000 chemical formulas of acids and acidic compounds, while in positive mode over 3000 unique elemental compositions of Nitrogen-Containing Aromatic Compounds were identified, illustrating the high accuracy of FT-ICR MS. The large data sets constituting FT-ICR MS spectra, require data treatment methods able to handle big data and find underlying relationships.

The objective of this review is to provide an overview of the machine learning methods used within the field of gas hydrates, with specific focus on the oil chemistry. First, we performed a text mining study to show the previous research areas of focus and expose potential gaps within. The aim of text mining is to scrape a web page of text related to a predefined keyword. We accessed all relevant publications from the Scopus Search database [41] and the most common and promising methods in literature are discussed. Additionally, methods commonly used for analysis of FT-ICR MS data in other fields which we believe could make valuable contributions to analysis of gas hydrate related samples, were identified. If correlations between hydrate-active components responsible for non-plugging crude oil systems and oil composition can be determined, this can be utilised as a parameter base for improved hydrate management strategies, better decision support tools and pipe flow simulations.

3. Text mining

To achieve an overview of the current status of machine learning methods within the field of petroleum gas hydrates the following questions were defined, of which the answers should give a thorough understanding of the field.

- **Q1**: Within which fields of gas hydrate research are machine learning used?
- **Q2**: What type of machine learning methods are used in the literature?
- **Q3**: What are the challenges in the field of gas hydrates using machine learning?
- **Q4**: How can machine learning improve the field of gas hydrate research?

3.1. Search strategy

For the text mining, we used the Scopus Search API from the pybliometrics library [42] in Python, which searches the Scopus database, containing over 78 million records within the fields of life sciences, social sciences, physical sciences and health sciences [43]. The search can be defined in different ways, searching for keywords, abstracts, title, doi, url, etc. Our approach was to search for selected words within either the keywords, titles or abstracts. To ensure that all relevant references were collected, the resulting literature was compared to results in Web of Science.

First a search was performed with the combination of *gas hydrates*, *FT-ICR MS*, *natural inhibitors* and *machine learning*, resulting in zero publications. The term *natural inhibitor* was removed and a search with *gas hydrates*, *FT-ICR MS* and *machine learning* was performed, which resulted in only one publication, a study performed by the authors of

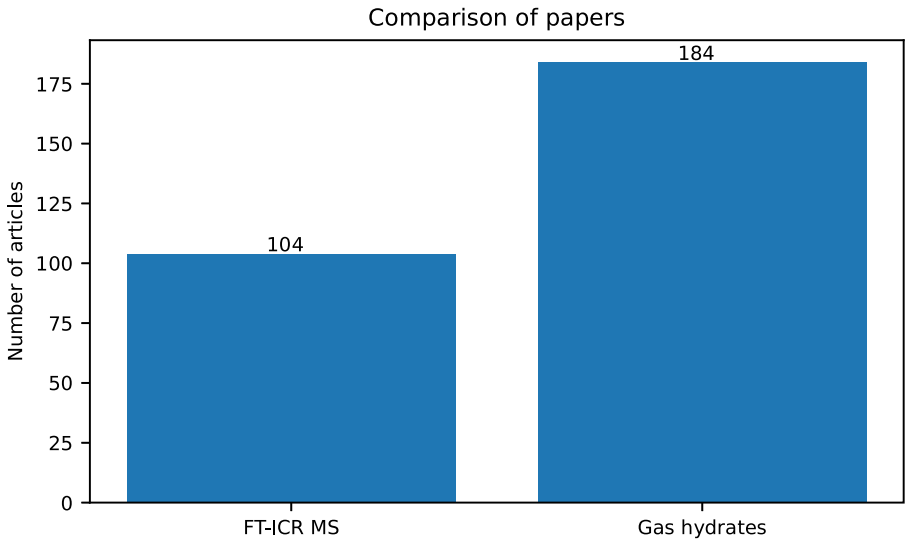

Fig. 1. Comparison of publications on the machine learning methods from Table 1 within the fields of gas hydrates and FT-ICR MS and number of publications retrieved.

Table 1
Overview of searches, each method was searched in combination with *gas hydrate* and *FT-ICR MS* to find all literature related to the methods.

Subjects	Methods
Gas hydrates FT-ICR MS	Principal Component Analysis (PCA) Partial Least Squares (PLS) Decision Trees (DT) Random Forest Artificial Neural Network (ANN) Support Vector Machine (SVM) Convolutional Neural Network (CNN) Regularisation/LASSO/Elastic Net/Ridge Regression Bayesian Networks (BN) K-Nearest neighbours (KNN)

this review [24]. Two new searches were therefore performed with *gas hydrates* plus *machine learning* and *FT-ICR MS* plus *machine learning*. This resulted in 45 publications for gas hydrates and 9 for FT-ICR MS. As very few publications were found, it was assumed that most publications do not use the term machine learning and only mention the methods used. Therefore, several machine learning methods were used as input in new searches. An overview of the methods included is presented in Table 1.

The resulting search phrases were as follows for gas hydrates *'TITLE-ABS-KEY((gas W/1 hydrate*) AND ((machine learning method) OR (method abbreviation)))'* and for FT-ICR MS *'TITLE-ABS-KEY((ft-icr W/1 ms) AND ((machine learning method) OR (method abbreviation)))'*. The 'W/1' ensures that the words are only one term apart and the * allows for different endings of the word, for instance *s* for plural notations. Duplicates of publications were removed.

A search was also performed for natural inhibitors with all the methods mentioned in Table 1 for both gas hydrates and FT-ICR MS, which resulted in zero publications.

To evaluate the use of mass spectrometry (MS) in the field of gas hydrates, a search was performed with *mass spectrometry* and *gas hydrates* which resulted in 2045 publications. To evaluate how many of these that were related to machine learning, a search with the methods presented in Table 1 was performed with both *mass spectrometry* and *gas hydrates*. This search resulted in 11 publications and all the 11 publications were also present in the results from the gas hydrate search with the machine learning methods.

The text mining study revealed that no other review paper exists on the topic of machine learning methods within the field of petroleum hydrates.

Text analysis was performed within the results of the two searches to find trends in the topics mentioned in the publications. The t-distributed stochastic neighbour embedding (t-SNE) technique was used to visualise the data. In t-SNE, similar data are grouped close together based on the stochastic neighbour embedding, while dissimilar data are more distant [44].

4. Results

The results from the two searches, *gas hydrates* and *FT-ICR MS*, with the methods in Table 1, are shown in Fig. 1. From the search of gas hydrates in combination with the methods from Table 1, 184 publications were retrieved and from FT-ICR MS and the methods in Table 1, 104 publications were retrieved. The publications returned by the text mining study are reported in the supporting information.

In Fig. 2 the publications on machine learning methods within the fields of gas hydrates and FT-ICR MS are plotted by publication year. Fig. 2 shows that there has been an increase in machine learning based research within both fields in the recent years. The first publication for gas hydrates was in 1998, and the first paper on use of machine learning within FT-ICR MS is from 2006. As FT-ICR MS has become more publicly available in the recent years, it is not surprising that the amount of publications have increased recently.

The amount of publications within each method is shown in Fig. 3. For gas hydrates, ANN is the most common machine learning method used, followed by SVM and PCA. For FT-ICR MS, the most common method is PCA followed by PLSR, the remaining methods have very few publications each and several of the methods had zero publications.

4.1. Text analysis study

A text analysis was performed, and a t-SNE plot of topics within the gas hydrate publications are shown with three topics in Fig. 4. The most common words for each topic are shown in the word clouds in Fig. 5. Fig. 4 shows that Topic 2 (orange) has the most entries of the three. The word clouds show that Topic 2 contains words such as 'gas', 'hydrate', 'prediction' and words associated with ANNs, 'artificial', 'neural' and 'network'. Topic 3 contains words associated with natural hydrates and some entries of 'network', while Topic 1 contains words associated with seismic and water analysis. From this analysis it is likely that the publications of interest with regards to machine learning and prediction of petroleum gas hydrates are within Topic 2 and natural gas hydrates within Topic 3.

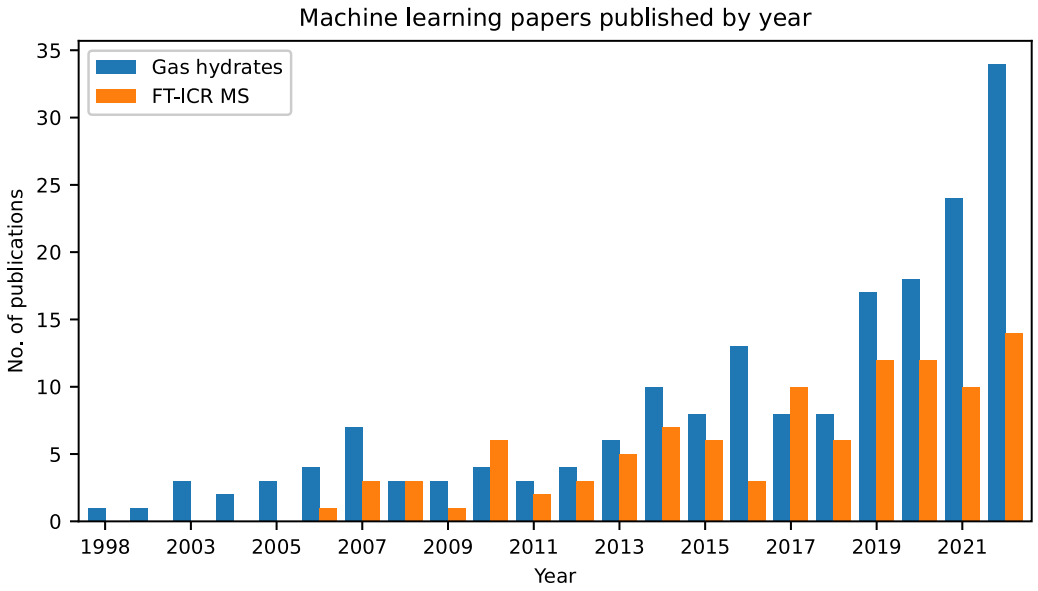

Fig. 2. The retrieved publications published by year, for gas hydrates in blue and FT-ICR MS in orange. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

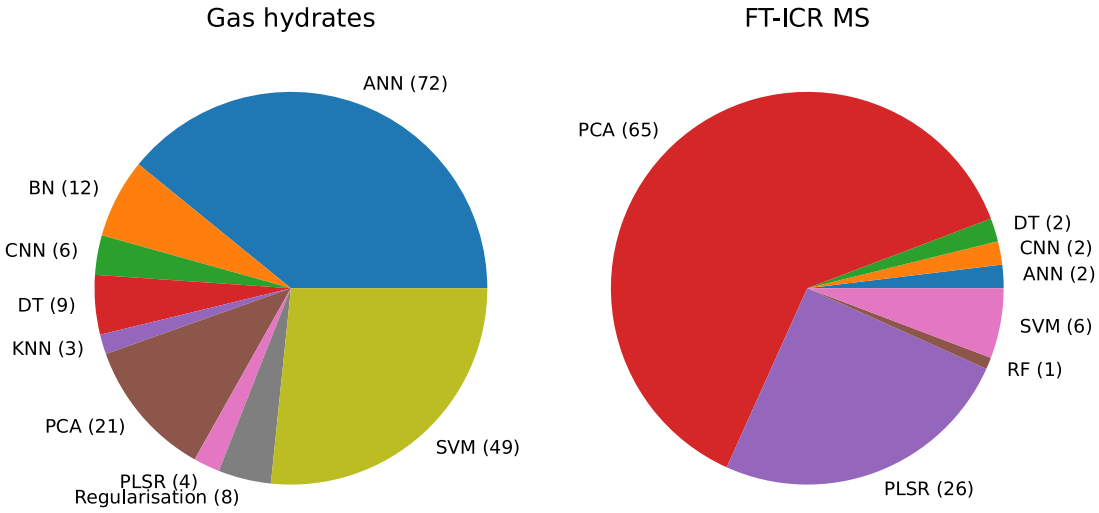

Fig. 3. Pie chart of the methods from Table 1 in combination with gas hydrates to the left and FT-ICR MS to the right with number of publications for each method in parenthesis.

A text analysis for the FT-ICR MS publications was also performed and the t-SNE plot with 3 topics is shown in Fig. 6. As t-SNE models similarities and dissimilarities, it is clear from Fig. 6 that Topic 1 (blue) is very different from Topic 2 (orange) as they are on the opposite sides of the plot, with Topic 3 (green) as a bridge between them. The most common words for each topic are shown in the word clouds in Fig. 7. Topic 1 is associated to oil spectroscopy and contains words from FT-ICR MS, Topic 2 contains words associated with organic matter analysis and Topic 3 contains metabolomic analyses. The machine learning studies performed on crude oils are therefore likely within Topic 1.

4.2. Classification vs. Regression

Machine learning can be used for analysis and visualisation of trends and allows identification of underlying phenomena in a data set. A typical pipeline for machine learning is displayed in Fig. 8. The process starts with collection of data, pre-processing, training of the model, testing of the model and finally deployment of the model through prediction from new data. The reader should seek out general textbooks for an introduction to machine learning [45,46].

Machine learning can be separated into two categories based on the desired response. When the response is continuous, regression analysis is used, while when the response is a discrete class label classification is used. Some algorithms can be used for both classification and regression tasks with only minor modifications. For gas hydrate purposes, both regression and classification methods are of interest. Which method to use is dependent on the type of data and the desired response to be predicted. For instance, when predicting thermodynamic properties of crude oils regression methods are most commonly used, as the desired prediction often is temperature, pressure or other measurements on the continuous scale. Classification methods are commonly used when samples are to be predicted based on their similarities to the defined classes. For instance when classifying oils into different types, properties etc.

In the following section, the methods included in the literature study and relevant references will be discussed to achieve an overview of the use of machine learning for analysis of petroleum related gas hydrates.

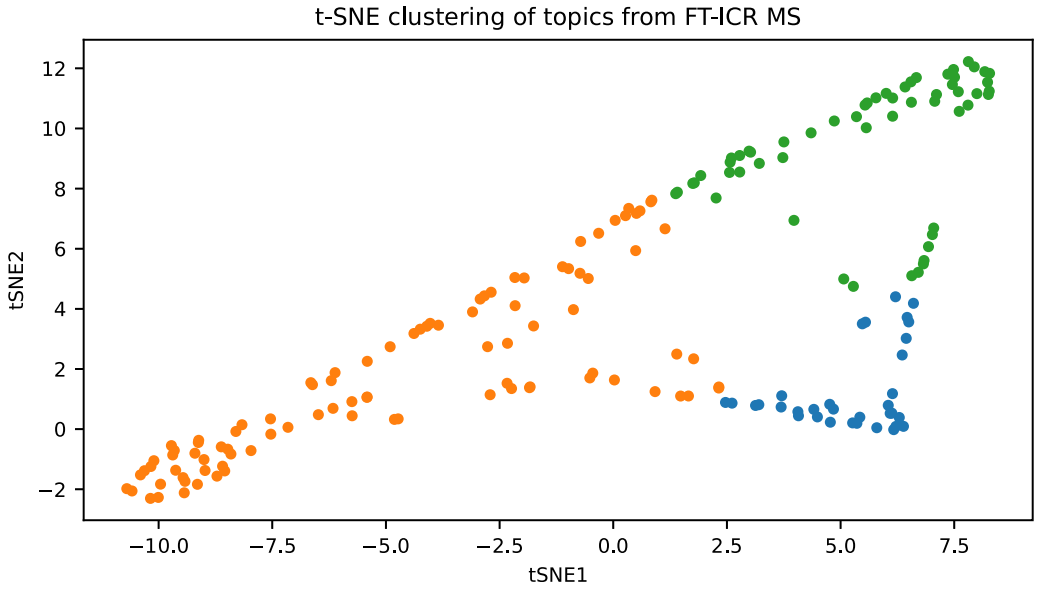

Fig. 4. t-SNE plot with three topics of the text analysis of machine learning publications on gas hydrates.

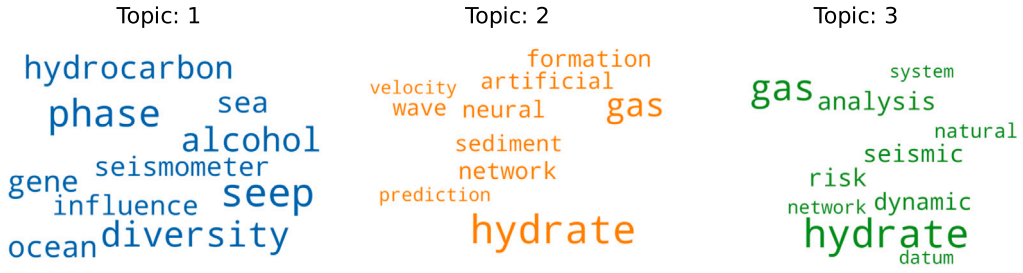

Fig. 5. Word clouds for each of the three topics and their most common words from the gas hydrate publications, with Topic 1 in blue, Topic 2 in orange and Topic 3 in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

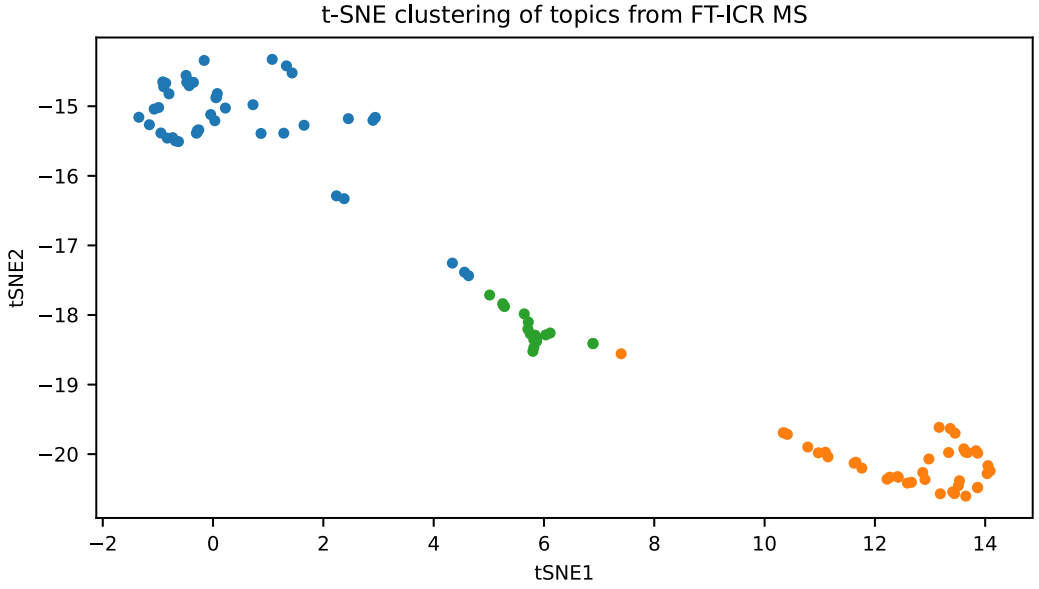

Fig. 6. t-SNE plot with three topics of the text analysis of machine learning publications on FT-ICR MS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

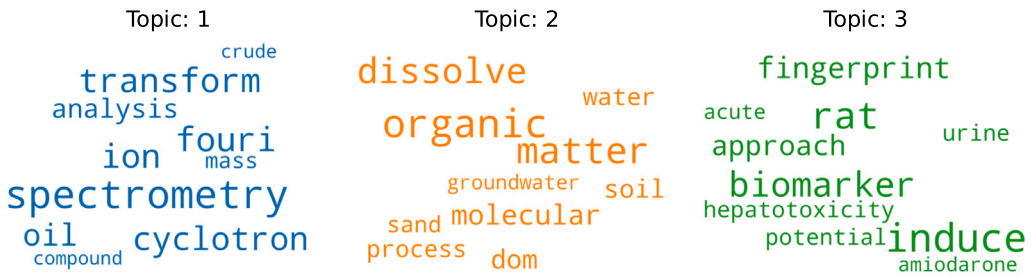

Fig. 7. Word clouds for each of the three topics and their most common words, from the FT-ICR MS publications, with Topic 1 in blue, Topic 2 in orange and Topic 3 in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

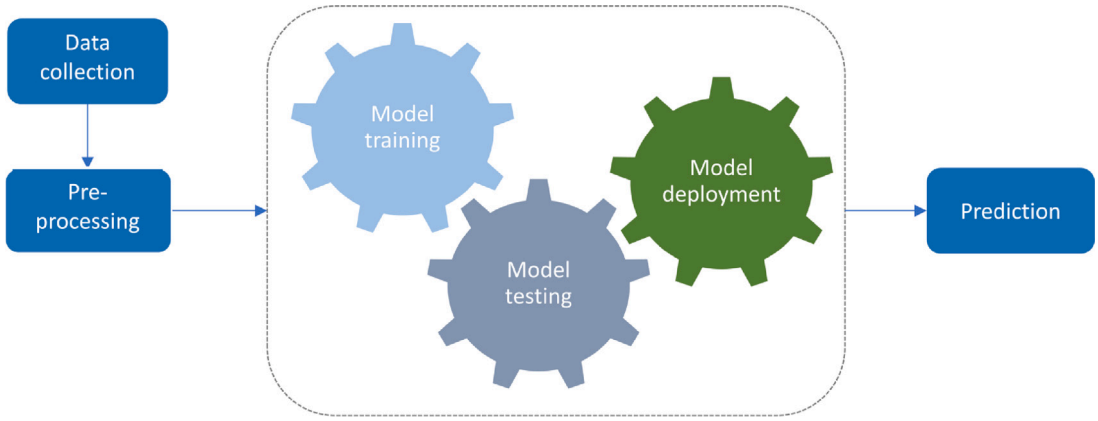

Fig. 8. Schematic illustration of a machine learning pipeline, with data collection, pre-processing, model training, testing, deployment and prediction.

4.3. Ordinary Least Squares (OLS)

OLS is a regression method for estimating the unknown parameters in a linear regression model. OLS minimises the sum of squares of the differences between the observed value and the value predicted by the linear function of the independent variable as shown by Eq. (1).

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \tag{1}$$

The coefficients ($\hat{\beta}$) can be estimated from Eq. (2).

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2}$$

A major drawback with OLS regression is that the matrix inversion used in the calculation of the regression coefficients requires the regressors to be linearly independent or uncorrelated. It also requires that the number of samples is larger than the number of variables, which is most often not the case when analysing data from FT-ICR MS. This renders OLS regression unsuitable for many data analysis problems. Two commonly used strategies, outlined below, to overcome this problem are (i) use of latent variables which represent linearly independent phenomena and (ii) regularisation.

4.4. Latent variable-based methods

4.4.1. Principal Component Analysis (PCA)

PCA [47] decomposes a large data set $\boldsymbol{X}$ into a subspace of latent variables representing the main features of variance as shown by Eq. (3).

$$\boldsymbol{X} = \boldsymbol{X}_{In} wgt_X \tag{3}$$

where $\boldsymbol{X}_{\mathrm{In}}$ is the data set with shape (N, K) for N samples and K variables, and wgt_{X} are the statistical weights balancing the sum of squares for the K $\boldsymbol{X}$-variables in $\boldsymbol{X}$, which has the shape (N, K). PCA is an effective dimension reduction technique that gives overview of large data sets and can be used prior to other data analysis methods in order to increase accuracy, overview and interpretation. Eq. (4) shows the PCA model for $\boldsymbol{A}$ Principal Components (PCs).

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A\boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{4}$$

where $\boldsymbol{P}_A$ are the loadings and orthonormal eigenvectors of $(\boldsymbol{X} - \bar{\boldsymbol{x}})^T(\boldsymbol{X} - \bar{\boldsymbol{x}})$ minimising the covariance between the $\boldsymbol{X}$-variables after $\boldsymbol{A}$ PCs. The scores ($\boldsymbol{T}_{\mathrm{A}}$) are orthogonal and calculated by Eq. (5).

$$\boldsymbol{T}_A = (\boldsymbol{X} - \boldsymbol{x})\boldsymbol{P}_A \tag{5}$$

The error term in Eq. (4) is $\boldsymbol{E}_A$ which is calculated by Eq. (6).

$$\boldsymbol{E}_A = \boldsymbol{X} - \bar{\boldsymbol{x}} - \boldsymbol{T}_A\boldsymbol{P}_A^T \tag{6}$$

PCA has commonly been used to identify correlations between analytical data and the properties of crude oils particularly from FT-ICR MS spectra as shown by the text mining study [48–52]. For instance, Hur et al. [49] analysed positive and negative mode APPI-FT-ICR MS spectra from 20 crude oils by PCA and identified differences between the oils based on their chemical composition. Moreover, their study showed a strong relationship between peaks in the mass spectra and the chemical properties of the oils indicating the potential for predicting crude oil properties from mass spectra.

4.4.2. Partial Least Squares Regression (PLSR)

PLSR [53] decomposes large data sets into a subspace of latent variables (scores and loadings) representing the main features of covariance between $\boldsymbol{X}$ (regressors) and $\boldsymbol{Y}$ (response). Both $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be multivariate. $\boldsymbol{X}$ has the same input model as for PCA shown in Eq. (3). As PLSR also takes the response into account, as opposed to PCA, there is an input model for $\boldsymbol{Y}$ which is shown in Eq. (7).

$$\boldsymbol{Y} = \boldsymbol{Y}_{In} wgt_Y \tag{7}$$

where $\boldsymbol{Y}_{\mathrm{In}}$ is the response with shape (N, J) for N samples and J response variables and wgt_{Y} are the statistical weights balancing the sum of squares for the J $\boldsymbol{Y}$-variables in $\boldsymbol{Y}$, which has the shape (N, J).

The decomposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is done simultaneously and iteratively, taking co-linearities in $\boldsymbol{Y}$ into account. For $\boldsymbol{X}$ the decomposition is shown in Eq. (8) and for $\boldsymbol{Y}$ in Eq. (9).

$$\boldsymbol{X} = \bar{\boldsymbol{x}} + \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A \tag{8}$$

$$\boldsymbol{Y} = \bar{\boldsymbol{y}} + \boldsymbol{U}_A \boldsymbol{Q}_A^T + \boldsymbol{F}_A \tag{9}$$

where $\boldsymbol{A}$ denotes the number of Principal Components (PCs) used and $\boldsymbol{E}_A$ and $\boldsymbol{F}_A$ are the error terms using $\boldsymbol{A}$ PCs. The loading weight matrix ($\boldsymbol{W}_A$) maximise the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by maximising the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$ with $\boldsymbol{A}$ PCs. The scores ($\boldsymbol{T}_A$) are orthogonal as shown by Eq. (10).

$$\boldsymbol{T}_A = (\boldsymbol{X} - \bar{\boldsymbol{x}})\boldsymbol{W}_A \tag{10}$$

The loadings for $\boldsymbol{X}$ ($\boldsymbol{P}_A$) are calculated by Eq. (11) while the loadings for $\boldsymbol{Y}$ ($\boldsymbol{Q}_A$) are calculated by Eq. (12).

$$\boldsymbol{P}_A = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{X} - \boldsymbol{x}) \tag{11}$$

$$\boldsymbol{Q}_A = (\boldsymbol{T}_A^T \boldsymbol{T}_A^T)^{-1} \boldsymbol{T}_A^T (\boldsymbol{Y} - \boldsymbol{y}) \tag{12}$$

The error term for $\boldsymbol{X}$ ($\boldsymbol{E}_A$) is calculated as for PCA in Eq. (6) and the error term for $\boldsymbol{Y}$ ($\boldsymbol{F}_A$) is calculated by Eq. (13).

$$\boldsymbol{F}_A = \boldsymbol{Y} - \bar{\boldsymbol{y}} - \boldsymbol{T}_A \boldsymbol{Q}_A^T \tag{13}$$

The regression coefficients ($\boldsymbol{B}_A$), which are measures of the impact of variations in the various regressors on the respective response variables, are calculated by Eq. (14).

$$\boldsymbol{B}_A = \boldsymbol{W}_A \boldsymbol{Q}_A^T \tag{14}$$

Prediction of $\boldsymbol{Y}$ for a new sample ($\boldsymbol{X}_{\text{new}}$) is then obtained by Eq. (15) where $\boldsymbol{b}_0$ is the intercept.

$$\boldsymbol{Y}_{pred} = \boldsymbol{b}_0 + \boldsymbol{X}_{new} \boldsymbol{B}_A + \boldsymbol{F}_A \tag{15}$$

PLSR has been widely used for analysis of mass spectra in a variety of application areas, including for gas hydrates and FT-ICR MS. Vaz et al. [51] correlated the chemical composition of crude oil from FT-ICR MS data with the total acid number (TAN), using PLSR and support vector machines (SVMs) as multivariate calibration methods. In Terra et al. [54] negative-ion mode electrospray ionisation, ESI(-)-FT-ICR MS was coupled to PLSR and variable selection methods to estimate the TAN of Brazilian crude oil samples. They showed that it was possible to relate the selected variables to their corresponding molecular formulas, thus identifying the main chemical species responsible for the TAN values. In Hemmingsen et al. [16] TAN values were also used as a response for PLSR to predict the acidic properties of the crude oils.

Terra et al. [55] predicted basic nitrogen and aromatics contents in crude oil, using positive ion mode laser desorption ionisation (LDI) coupled to FT-ICR MS and PLSR with variable selection based on competitive adaptive reweighted sampling (CARS) in a procedure called CARSPLS regression.

Lozano et al. [56] used PLSR and genetic algorithm variable selection on APPI(+)-FT-ICR MS data for quantitative analysis of crude oils and their fractions. They estimated the API gravity and Conradson Carbon Residue of Colombian crude oil and vacuum residue (VR) samples with high accuracy.

PLSR can also be used for classification problems, for instance in the combination with discriminant analysis (DA), as PLS-DA. Two common DA methods are Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) which model the class conditional distribution of the data $P(X|y = k)$ for each class k. Predictions are obtained by using Bayes' rule, and the class that maximises this conditional probability is selected. The class priors $P(y = k)$ (the proportion of instances of class k), the class means and the covariance matrices are then estimated from the training data.

In Chua et al. [57] PLS-DA was used in tandem with PCA to analyse crude oil spill data from gas chromatography techniques. The PLS-DA and PCA combination accurately characterised the crude oil spill samples, overcoming the shortcomings of the traditional methods.

Likewise, Melendez-Perez et al. [58] utilised PLS-DA for analysis of ESI(-)-FT-ICR MS spectra of lacustrine oil and marine oil samples aiming towards comparing and classifying the samples. Results show that FT-ICR MS coupled with PLS-DA has potential to reveal oil characteristics more clearly.

Gjelsvik et al. [24] was the only publication from the text mining results regarding natural inhibitors. In this study, machine learning-based variable selection was used to identify components related to gas hydrate formation and PLS-DA emerged as the best performing method. This study showed that it is possible to identify features from FT-ICR MS spectra related to hydrate formation.

Accordingly, PLSR have already been shown to be able to predict chemical properties of crude oils, and PLS-DA has been shown to be able to classify crude oils samples with high accuracy.

4.4.3. Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR)

One promising extension of PLSR is the HC-PLSR [59] method, which is a locally linear regression method based on separating the observations into clusters and generating local PLSR models within each cluster. A global PLSR model comprising all observations is first made, and the observations are clustered based on the scores from this PLSR model. Local PLSR models are then made within each cluster. New observations are projected into the global model and classified based on their predicted $\boldsymbol{X}$-scores. Prediction of the response is based on either the closest local model or a weighted sum of all local models. HC-PLSR can be used with any clustering and classification method. HC-PLSR allows for local analysis within each cluster, and represents a way to handle highly nonlinear relationships between the regressors and the response.

4.4.4. Artificial Neural Networks (ANNs)

ANNs [60–62] are computing systems consisting of nodes called artificial neurons, between which the connections have numeric weights that are often initialised at random, and adjusted by backpropagation. Backpropagation uses the prediction error to calculate the gradient of the loss function with respect to the weights in the network. The neurons are placed in different layers, typically an input layer, one or more hidden layers, and an output layer. A widely used type of composition is the nonlinear weighted sum given by Eq. (16).

$$f(x) = K\left(\sum_i w_i g_i(x)\right) \tag{16}$$

where K is the activation function (some predefined function, such as the hyperbolic tangent or a sigmoid function), w_i are the weights and g_i are the different functions that are combined in the network. As ANNs use self learning, the network can adjust weights when a new situation is introduced, which leads to more flexible predictions than traditional regression models. ANNs are trained with experimental data where the output is a nonlinear function of the input data after learning a pattern and creating a prediction model [63]. Deep Neural Networks [64] are ANNs with multiple hidden layers between the input and output layers, as shown in Fig. 9. These can contain many layers of nonlinear hidden units.

Elgibaly and Elkamel [65,66] were the first to develop ANNs to predict thermodynamic conditions and suitable inhibitors for gas hydrate systems. Their network performed well compared to previous prediction methods based on traditional statistics and experimental data analysis, but showed signs of overfitting supposedly due to lack of experimental data. Chapoy et al. [67] used feed-forward neural networks (FNNs) to predict hydrate stability zones achieving a reasonable

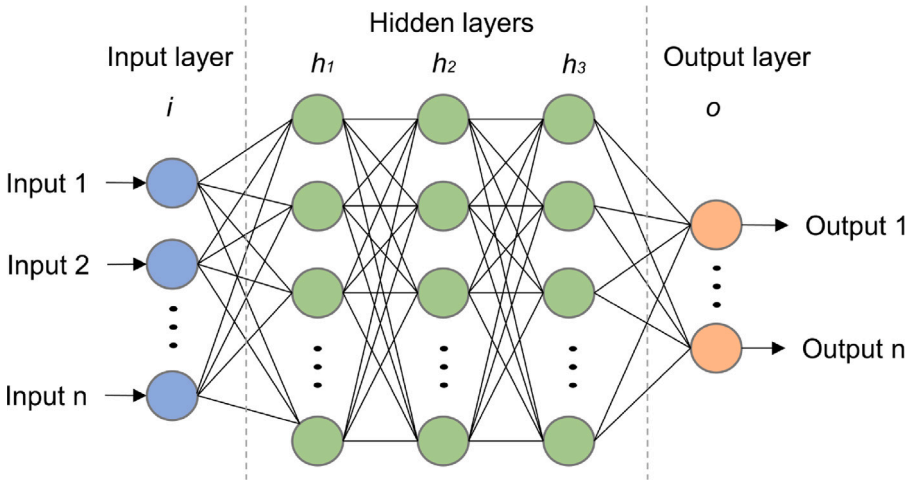

Fig. 9. Schematic example of a neural network with input layer, hidden layers and output layer.

model, but also pointing out deficiencies in the experimental data as a weakness of the study.

Ghavipour et al. [68] constructed an apparatus that measured specific gravity of different gas mixtures and pressure during a hydrate formation process. ANNs were then used to predict the hydrate formation conditions by a network with two hidden layers and 10 neurons in each layer, validated with Leave-One-Out cross validation.

Several studies have in the recent years used ANNs to predict hydrate formation conditions [69–71]. The purpose of these types of predictions is to identify the conditions where gas hydrates are formed and avoid operation within this region.

4.4.5. Support Vector Machines (SVMs)

SVMs [72] are supervised learning methods that analyse data for classification or regression analysis. SVMs are well suited for learning tasks where the number of variables is large compared to the number of observations in the training set.

For classification, SVMs construct a hyperplane or a set of hyperplanes in a high-dimensional space to separate the observations into two groups [73]. The goal is to find the hyperplane that has the largest distance (margin) to the nearest data point belonging to any of the two classes. The margin is defined as the distance between the separating hyperplane (decision boundary) and the training samples that are closest to this hyperplane. Data points that lie on the margin are known as support vector points, and the solution is represented as a linear combination of only these points. Decision boundaries with large margins tend to have a lower generalisation error, while decision boundaries with small margins are more prone to overfitting.

SVMs can be applied to nonlinear classification problems by using the so-called kernel trick, where the original space is mapped into a much higher-dimensional space where the observations can be more easily separated. To achieve this, a mapping function φ is used, as shown in Fig. 10. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant.

In Support Vector Regression (SVR), the hyperplane is the line that is used to predict the continuous output, shown in Fig. 11. SVR basically considers the points that are within the decision boundary lines and the regression line is then the hyperplane that has a maximum number of points.

SVMs are the second most commonly used methods for gas hydrates. Cao et al. [74] developed an SVM model for predictions of gas hydrate formation conditions, in combination with selection algorithms to optimise the process parameters for the SVM. Qin et al. [75] used both SVM and ANNs to predict gas hydrate plugging risks from flowloop and field data with SVM outperforming the ANN.

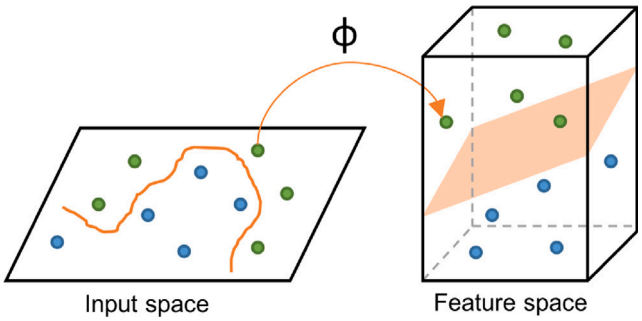

Fig. 10. The kernel trick to handle non-linear problems.

Rashid et al. [76], Mesbah et al. [77], Ghiasi et al. [78] and Yarveicy and Ghiasi [79] created SVM models with a linear modification of the SVM algorithm known as the least squares support vector machine (LSSVM) to predict thermodynamic properties of gas hydrate systems. One drawback with SVMs is the large number of quadratic computations performed to analyse the data, requiring high computational power, but LSSVM overcomes this due to the less complicated calculation methods [80].

As previously mentioned, Vaz et al. [51] predicted the TAN from FT-ICR MS spectra with SVM performing better than both PLSR and univariate methods. SVM is thereby able to both predict thermodynamic properties of hydrates and chemical properties of crude oils.

4.4.6. Decision Trees (DTs)

DTs [81,82] are attractive models when interpretability is important, and consist of a tree root, internal nodes, branches and leaf nodes. DTs ask a series of questions, and generate decision rules based on these. The model seeks to find the smallest set of rules that is consistent with the training data. In general, the rules have the form: *if* $condition_1$ *and* $condition_2$ *and* $condition_3$ *then outcome*. Fig. 12 shows an illustration of a decision tree model.

The rules are chosen to divide observations into segments that have the largest difference with respect to the target variable. Thus the rule selects both the variable and the best break point to separate the resulting subgroups maximally. The break points of variables are found using significance testing (F- or Chi-square with Bonferroni corrections) or reduction in variance criteria. To avoid overfitting, one often has to prune the tree by setting a limit for the maximal depth of the tree. A leaf can no longer be split when there are too few observations, the maximum depth (hierarchy of the tree) has been reached, or

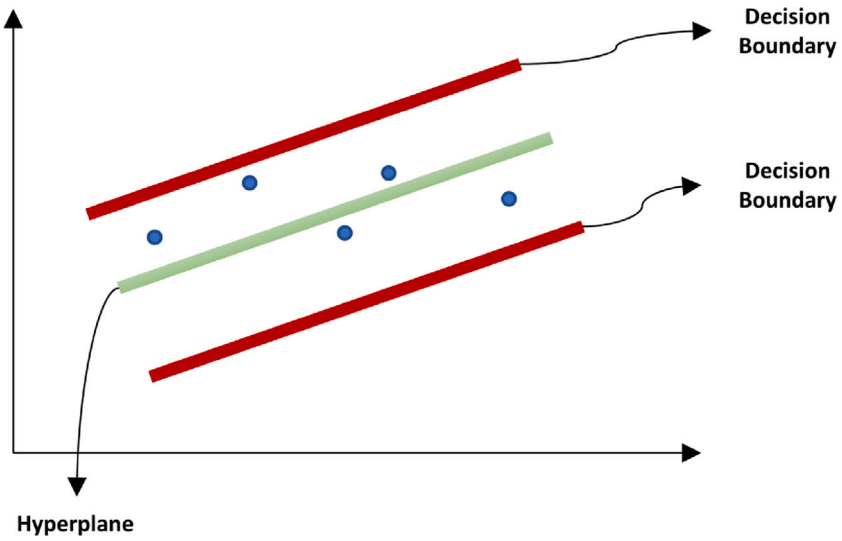

Fig. 11. Illustration of the hyperplane and decision boundaries in SVR.

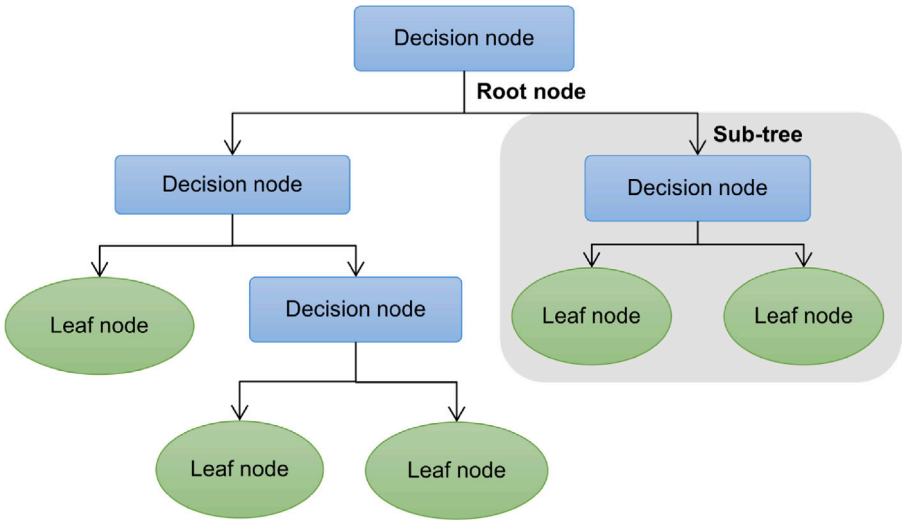

Fig. 12. Illustration of decision trees with the root node, sub-trees, decision nodes, branches and leaf nodes.

no significant split can be identified. It is assumed that observations belonging to different classes have different values in at least one of their variables. DTs are usually univariate, since they use splits based on a single feature at each internal node, but methods are available for constructing multivariate trees [83].

To improve the prediction of the DT, a boosting method can be applied. Boosting is an ensemble method for improving predictions of a weak learning algorithm [84]. The weak learners are trained sequentially, trying to improve upon its predecessor. When boosting is applied to a tree, each tree is dependent on prior trees and the algorithm learns by fitting the residual of the prior trees. One example of a boosting method is XGBoost (eXtreme Gradient Boosting). In XGBoost, trees are built at every iteration, always minimising the prediction error of the classifier while introducing a penalty function to utilise the computational power more efficiently.

4.4.7. Random Forest (RF)

In DTs, the initial selected split affects the optimality of variables considered for subsequent splits. Ensemble tree models grow trees with varying initial splits, and use either a voting or the average of the predictions for each new data point across all trees. The vote distribution can be used to develop a nonparametric probabilistic predictive model. The ensemble is less prone to overfitting and other problems of individual DTs, and generally performs better. RF [85–87] is an example of such an ensemble tree method. For RF, each tree is based on a random subset of the data and variables (selected by bootstrapping). The change in prediction accuracy when the values of a feature are randomly permuted among observations gives estimates of the importance of each feature.

Tree models and boosting are among the most common regression and classification methods, and has been used for gas hydrates and crude oil analysis. Song et al. [88] used a gradient boosted regression tree algorithm to predict hydrate phase equilibrium conditions in the presence of various salts, organic substances or water. The model was compared to an ANN, where the regression tree achieved the best prediction model for gas hydrates' phase equilibrium conditions in the presence of various salts or organics.

In Acharya and Bahadur [89] RF and XGBoost were used to predict gas hydrate dissociation temperatures in the presence of hydrate inhibitors and precursors achieving good predictions.

Lovatti et al. [90] proposed two strategies for the use of RF and data reduction techniques for NMR spectra of petroleum samples. The study compared the NMR spectra to the TAN values of the petroleum, and

the method was able to identify a relationship between the TAN and specific regions in the spectra.

4.4.8. Naive Bayes (NB) classification

A Bayesian network (BN) is a probabilistic model that represents a set of random variables and their conditional independence via a directed acyclic graph (DAG). Using e.g. Chi-squared and mutual information tests, one can find the conditional independence relationships among the variables and use these relationships as constraints to construct a BN. BNs can take prior knowledge into account, by e.g. setting a certain node as *root node* or *leaf node*, thereby applying knowledge of nodes that are direct causes or effects of other nodes. This results in nodes that are not directly connected to another node, or that two nodes are independent.

The probabilistic parameters are encoded into a set of tables, one for each variable, in the form of local conditional distributions of a variable given its parents. The joint distribution can be reconstructed by multiplying these tables (given the independencies encoded into the network). BNs are DAGs whose nodes represent random variables that may be e.g. observable quantities or latent variables. Edges (connections) represent conditional dependencies, and each node is associated with a probability function.

Naïve Bayesian networks are very simple BNs which are composed of DAGs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes), where the child nodes are assumed to be independent. Naïve Bayes (NB) classification may be impaired by probabilities of 0, but this can be avoided by using a Laplace estimator.

The assumption of independence among child nodes is most often not valid, but this can be corrected for by adding extra edges to include some of the dependencies between the variables. In this case, the network has the limitation that each feature can be related to only one other feature [91]. Selective Bayesian classifiers [92] include a feature selection stage to remove irrelevant variables or one of the two totally correlated variables.

Shi et al. [93] used a variational Bayesian neural network for probabilistic deepwater natural hydrate gas dispersion modelling of simulated data. Combined with a convolutional neural network, the model performed well.

Bayesian networks have been used for risk and safety assessment of storing and transportation of crude and heavy oil. For example, Zhang et al. [94] used BNs to evaluate the leak safety of heavy oil gatherings in pipelines. BNs find the probability for leakage and fuzzy set theory evaluates the consequences of the leakage.

4.4.9. k-nearest neighbours (KNN) classification

KNN [95] locates the k nearest observations to the observation to be classified (e.g. by an exhaustive search algorithm) based on the chosen distance metric, and identifies the most frequent class membership among the neighbours. The number k is specified by the user, and the right choice of k is crucial to find a good balance between overfitting and underfitting. Weights are assigned to the contributions of the neighbours in a majority voting to predict the classes, so that the nearer neighbours contribute more to the average than the more distant ones.

KNN is fundamentally different from the other supervised classifiers described here, in that it is a so-called lazy learner. KNN does not learn a discriminative function from the training data but memorises it instead. The main advantage of such a memory-based approach is that the classifier immediately adapts as we collect new training data. However, the computational complexity for classifying new samples grows with the number of samples in the training data set and storage space can hence become a challenge when working with large data sets.

Only two instances where KNN were used related to gas hydrates were found. Xu et al. [96] used KNN regression, SVM, RF and XGBoost for the prediction of hydrate formation temperatures achieving good predictions with all methods.

Amin et al. [97] used KNN to predict hydrate equilibrium conditions to CO_2 capture. The model was simple but showed good predictions with low errors, indicating that KNN is a valuable method for analysis of gas hydrate thermodynamics.

4.5. Regularisation-based methods

Another group of machine learning methods that we find promising for identification of hydrate active compounds in crude oils is regularisation-based methods, which are very useful for feature selection purposes. The most commonly used regularisation-based methods are Ridge regression [98], LASSO (least absolute shrinkage and selection operator) [99] and Elastic net [100]. Regularisation-based versions of PLSR are also available, which have shown promise in feature selection, such as Sparse-PLS [101]/Soft-Threshold PLS [102] and Powered PLS [103]. These may have advances over other regularisation-based methods in cases where interpretation is important, due to the possibilities to gain overview of complex data sets through decomposition of the data into a lower-dimensional subspace of latent variables.

4.5.1. Ridge regression

Ridge regression is also known as L2-regularisation. In Ridge, the sum of the squares of the regression coefficients (β) is forced to be less than a fixed value, which shrinks the size of the coefficients. Ordinary least squares (OLS) minimises Eq. (17).

$$RSS_{OLS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \tag{17}$$

while Ridge regression minimise Eq. (18).

$$RSS_{Ridge} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{i=1}^{p}\beta_j^2 \tag{18}$$

where $\lambda \geq 0$ is a penalty term which is often found by cross-validation. This gives Eqs. (19) and (20).

$$B_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{19}$$

$$B_{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{20}$$

Hence, Ridge regression handles multicollinearity in the regressor ($\boldsymbol{X}$) matrix, while OLS regression does not.

4.5.2. LASSO

In LASSO, the estimates of the regression coefficients are obtained using L1-constrained least squares. This forces the sum of the absolute values of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero. LASSO is a feature selection method, since variables having zero regression coefficients are omitted from the model. In LASSO Eq. (21) is minimised.

$$RSS_{LASSO} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{i=1}^{p}\beta_j \tag{21}$$

4.5.3. Elastic net

Elastic net combines the L1 and L2 penalties of the Ridge and LASSO methods linearly as given by Eq. (22).

$$RSS_{EN} = \sum_{i=1}^{n}\left(y_i - \beta_0\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{i=1}^{p}\beta_j^2 + \lambda_2\sum_{i=1}^{p}\beta_j \tag{22}$$

In Elastic net, highly correlated regressors will tend to have similar regression coefficients, which creates a grouping effect that is desirable in many applications.

Landgrebe and Nkazi [104] used traditional L1/L2 in order to reduce overfitting of the neural network, but dropout regularisation proved more effective.

In Singh et al. [105] Ridge Regression (L2) was used among other methods to estimate gas hydrate saturation in sedimentary systems from well-logs by NMR measurements. L2 achieved good accuracy and was one of the best performing methods. This is an indication that L2 could also perform well with other spectroscopic data of gas hydrate related samples, such as FT-ICR MS spectra.

Similarly, other regularisation methods have been used in combination with spectroscopic data previously. In Fu et al. [106] both Sparse-PLS and Elastic net were used for wavelength selection on data from NIR spectroscopy of corn and gasoline. Both methods select intervals of wavelengths, where Elastic net selects a smaller model, while Sparse-PLS achieves a higher accuracy. Finding the wavelengths closely related to the response could significantly improve a prediction model.

4.5.4. Convolutional Neural Networks (CNNs)

CNNs are deep neural networks which use convolutions to extract information in one or more of the hidden layers [63]. CNNs are regularised versions of fully connected networks. In a convolutional layer, the data is organised in a feature map where the weights are connected to the previous layer. These weights are used to filter for patterns in the data. Commonly used in pattern recognition, CNNs are good feature extractors by learning the most important variables by itself.

CNNs can be a valuable tool for instance for the analysis of mass spectrometry data. Lv et al. [107] used CNNs to analyse peak information in tandem mass spectrometry (MS/MS). This method outperformed others such as SVMs, PCA, deep neural networks and XGBoost. Due to the nature of the convolutional filters, CNNs are able to learn both the peak shape and the m/z values, achieve greater robustness for low signal-to-noise ratios and can allow for a higher-level representation of lower-level features representing patterns [108]. Hence, CNNs could be very useful for analysing FT-ICR MS data.

Kim et al. [109] used CNNs for saturation modelling from X-ray CT images. The 1-dimensional CNNs performed well, but the method shows difficulties in determination of optimal parameters for the CNNs.

Li et al. [110] constructed a neural network based on a variational autoencoder with convolutional layers to predict pore size distributions in subsurface shale reservoirs. The method showed good predictions and although this is not directly related to gas hydrates, gas hydrates are analysed in a similar manner, indicating that CNN could be a valuable method given an optimal parameter search.

4.6. Data used in literature

The data used in many of the machine learning models previously developed in the field of gas hydrates have been sampled from the literature. In this review, a number of the cited articles discussed are based on data sampled from other publications [65–67,69–71,74,77,78,88,89,96,104]. These references are mainly based on thermodynamic data, concerning prediction of gas hydrate formation/dissociation conditions and phase equilibrium measurements. Sloan and Koh [1] present an extensive list of experimental data which are frequently used by the authors sampling experimental data from the literature [65–67,69,71,104]. Consequently, the models from these authors are based on the same data. This can result in shortcomings, as the errors in predictions from these models approximate the errors of the experiments. Additionally, where the data are deficient, extrapolation has to be performed which decreases the accuracy of the predictions [3]. It is therefore clear that there is a need for more experimental data. New experimental data should fill the gaps in already published data, and as many of the models are based on thermodynamic properties, other aspects of gas hydrates could be valuable to examine closer. Better understanding of the mechanisms and the molecular composition related to the inhibition/dissociation of gas hydrates, could lead to strengthened prediction models for the thermodynamic, physical and chemical properties of gas hydrates in the future.

5. Conclusions and future perspectives

In this paper a text mining study was performed to evaluate the use of machine learning methods within the field of gas hydrates with specific focus on the oil chemistry. An evaluation of FT-ICR MS was included in the study to establish a link between aspects of gas hydrates and analysis of crude oils. Several machine learning methods were identified as promising and their use in the literature was evaluated. For studies regarding gas hydrates, predictions of thermodynamic properties were most common, while FT-ICR MS was used for analysis of oil chemistry and chemical properties. Most of the publications on thermodynamic properties of gas hydrates were also created using the same data sources. It could therefore be beneficial to explore other areas of gas hydrate research using machine learning in the future. Although there is little literature describing the use of FT-ICR MS to characterise gas hydrates, the text mining results show that FT-ICR MS has been used to characterise crude oils for some time and with success. Therefore, with the combination of FT-ICR MS and machine learning, it may be possible to identify the hydrate-active compounds responsible for the differences between oils forming plugging hydrates and oils forming transportable hydrates. This can be done by relating the composition of the oil, determined by FT-ICR MS to information regarding hydrate formation. The methods presented in this paper successfully predicted thermodynamic properties in gas hydrates or chemical properties from FT-ICR MS, and the methods could therefore be tested with the aim of predicting chemical properties from gas hydrate related samples. We believe that an approach which is able to predict hydrate behaviour may lead to new knowledge about natural gas hydrate inhibitors. The development of a universal method to identify natural components which inhibit, or work as AAs for gas hydrates would contribute to new understanding and decision making tools in the field of gas hydrate flow assurance and management strategies. This could lead to better decision support tools and better risk evaluations for transportation of crude oils with gas hydrates present.

The text mining study revealed that the amount of research using machine learning to analyse both gas hydrate and FT-ICR MS data is still limited, but research on both topics have increased in recent years. For FT-ICR MS, most publications used PCA for analysis of the data, and several of the publications used the chemical composition data to build machine learning models instead of using the mass spectra directly. Identifying relationships and building models based on the mass spectra requires less pre-processing steps and could therefore be advantageous and could be explored further.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The authors thank the The Research Council of Norway, Equinor ASA, Norway, OMV (Norge) AS, Norway, Wintershall DEA Norge AS and TotalEnergies, Norway for funding. This work is a part of the Knowledge-Building Project for Industry (PETROMAKS 2), Project number: 294636 ‘‘New Hydrate Management: New understanding of hydrate phenomena in oil systems to enable safe operation within the hydrate zone’’.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.fuel.2022.126696.**S1 Table. Results from the text mining study**. Table of the publications returned by the text mining study.

References

[1] Sloan ED, Koh CA. Clathrate hydrates of natural gases. Chemical industries series, 3rd ed.. vol. 119, Boca Raton, FL: CRC Press, Taylor & Francis Group; 2008.

[2] Fotland P, Askvik KM. Some aspects of hydrate formation and wetting. J Colloid Interface Sci 2008;321:130–41.

[3] Sloan ED. A changing hydrate paradigm—from apprehensionto avoidance to risk management. Fluid Phase Equilib 2005;228–229:67–74.

[4] Kelland MA. History of the development of low dosage hydrate inhibitors. Energy Fuels 2006;20:825–47.

[5] Nasir Q, Suleman H, Elsheikh YA. A review on the role and impact of various additives as promoters/ inhibitors for gas hydrate formation. J Nat Gas Sci Eng 2020;76:103211.

[6] Sa J-H, Melchuna A, Zhang X, Rivero M, Glénat P, Sum AK. Investigating the effectiveness of anti-agglomerants in gas hydrates and iceformation. Fuel 2019;255:115841.

[7] Ding L, Shi B, Liu Y, Song S, Wang W, Wu H, Gong J. Rheology of natural gas hydrate slurry: Effect of hydrate agglomeration anddeposition. Fuel 2019;239:126–37.

[8] Lederhos J, Longs J, Sum A, Christiansen RL, Sloan Jr ED. Effective kinetic inhibitors for natural gas hydrates. Chem Eng Sci 1995;51:1221–9.

[9] Shahnazar S, Bagheri S, TermehYousefi A, Mehrmashhadi J, Karim MSA, Kadri NA. Structure, mechanism, and performance evaluation of natural gas hydrate kinetic inhibitors. Rev Inorg Chem 2018;38:1–19.

[10] Lingelem MN, Majeed AI, Stange E. Industrial experience in evaluation of hydrate formation, inhibition, and dissociation in pipeline design and operation. Ann New York Acad Sci 1994;715:75–93.

[11] Fadnes FH. Natural hydrate inhibiting components in crude oils. Fluid Phase Equilib 1996;117:186–92.

[12] Borgund AE, Høiland S, Barth T, Fotland P, Askvik KM. Molecular analysis of petroleum derived compounds that adsorbonto gas hydrate surfaces. Appl Geochem 2009;24:777–86.

[13] Høiland S, Askvik KM, Fotland P, Alagic E, Barth T, Fadnes F. Wettability of Freon hydrates in crude oil/brine emulsions. J Colloid Interface Sci 2005;287:217–25.

[14] Høiland S, Borglund AE, Barth T, Fotland P, Askvik KM. Wettability of Freon hydrates in crude oil/brine emulsions: the effects of chemical additives. In: 5th international conference in gas hydrate, Vol. 4. Trondheim; 2005, p. 1151–61.

[15] Borgund AE, Erstad K, Barth T. Fractionation of crude oil acids by HPLC and characterization of their properties and effects on gas hydrate surfaces. Energy Fuels 2007;21:2816–26.

[16] Hemmingsen PV, Kim S, Pettersen HE, Rodgers RP, Sjöblom J, Marshall AG. Structural characterization and interfacial behavior of acidic compounds extracted from a North Sea oil. Energy Fuels 2006;20:1980–7.

[17] Hemmingsen PV, Li X, Peytavy J-L, Sjöblom J. Hydrate plugging potential of original and modified crude oils. J Dispers Sci Technol 2007;28:371–82.

[18] Erstad K, Høiland S, Fotland P, Barth T. Influence of petroleum acids on gas hydrate wettability. Energy Fuels 2009;23:2213–9.

[19] Qiao P, Harbottle D, Tchoukov P, Masliyah J, Sjoblom J, Liu Q, Xu Z. Fractionation of asphaltenes in understanding their role in petroleum emulsion stability and fouling. Energy Fuels 2016;31:3330–7.

[20] Salmin DC. The impact of synthetic and natural surface-active components on hydrate agglomeration (Doctoral thesis), Golden, Colorado: Colorado School of Mines; 2019.

[21] Adams JJ. Asphaltene adsorption, a literature review. Energy Fuels 2014;28:2831–56.

[22] Kilpatrick PK. Water-in-crude oil emulsion stabilization: Review and unanswered questions. Energy Fuels 2012;26:4017–26.

[23] Yang F, Tchoukov P, Dettman H, Teklebrhan RB, Liu L, Dabros T, Czarnecki J, Masliyah J, Xu Z. Asphaltene subfractions responsible for stabilizing water-in-crude oil emulsions. Part 2: Molecular representations and molecular dynamics simulations. Energy Fuels 2015;29:4783–94.

[24] Gjelsvik EL, Fossen M, Brunsvik A, Tøndel K. Using machine learning-based variable selection to identify hydrate related components from FT-ICR MS spectra. PLoS One 2022;17(8):e0273084.

[25] Marshall AG, Rogers RP. Petroleomics: The next grand challenge for chemical analysis. Acc Chem Res 2004;37:53–9.

[26] Hughey CA, Rodgers RP, Marshall AG. Resolution of 11 000 compositionally distinct components in a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of crude oil. Anal Chem 2002;74:4145–9.

[27] Cho Y, Ahmed A, Islam A, Kim Sunghwan. Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. Mass Spectrom Rev 2014;34:248–63.

[28] Emmett MR, White FM, Hendrickson CL, Shi SD-H, Marshall AG. Application of micro-electrospray liquid chromatography techniques to FT-ICR MS to enable high-sensitivity biological analysis. J Am Soc Mass Spectrom 1998;9:333–40.

[29] Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG. Kendrick mass defect spectrum: A compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal Chem 2001;73:4676–81.

[30] Marshall AG, Rodgers RP. Petroleomics: Chemistry of the underworld. Proc Natl Acad Sci USA 2008;105:18090–5.

[31] de Hoffmann E, Stroobant V. Mass spectrometry: Principles and applications. 3rd ed.. West Sussex, England: John Wiley and Sons Ltd.; 2012.

[32] Hur M, Yeo I, Kim E, No M-h, Koh J, Cho YJ, Lee JW, Kim S. Correlation of FT-ICR mass spectra with the chemical and physical properties of associated crude oils. Energy Fuels 2010;24:5524–32.

[33] Klein GC, Kim S, Rodgers RP, Marshall AG, Yen A. Mass spectral analysis of asphaltenes. II. Detailed compositional comparison of asphaltenes deposit to its crude oil counterpart for two geographically different crude oils by ESI FT-ICR MS. Energy Fuels 2006;20:1973–9.

[34] Schaub TM, Jennings DW, Kim S, Rodgers RP, Marshall AG. Heat-exchanger deposits in an inverted steam-assisted gravity drainage operation. Part 2. Organic acid analysis by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Energy Fuels 2007;21:185–94.

[35] Smith DF, Rahimi P, Teclemariam A, Rodger RP, Marshall AG. Characterization of athabasca bitumen heavy vacuum gas oil distillation cuts by negative/positive electrospray ionization and automated liquid injection field desorption ionization Fourier transform ion cyclotron resonance mass spectrometry. Energy Fuels 2008;22:3118–25.

[36] Headley JV, Peru KM, Barrow MP, Derrick PJ. Characterization of naphthenic acids from athabasca oil sands using electrospray ionization: The significant influence of solvents. Anal Chem 2007;79:6222–9.

[37] Barrow MP, Headley JV, Peru KM, Derrick PJ. Data visualization for the characterization of naphthenic acids within petroleum samples. Energy Fuels 2009;23:2592–9.

[38] Fernandez-Lima FA, Becker C, McKenna AM, Rodgers RP, Marshall AG, Russell DH. Petroleum crude oil characterization by IMS-MS and FTICR MS. Anal Chem 2009;81:9941–7.

[39] Qian K, Robbins WK, Hughey CA, Cooper HJ, Rodgers RP, Marshall AG. Resolution and identification of elemental compositions for more than 3000 crude acids in heavy petroleum by negative-ion microelectrospray high-field Fourier transform ion cyclotron resonance mass spectrometry. Energy Fuels 2001;15:1505–11.

[40] Qian K, Rodgers RP, Hendrickson CL, Emmett MR, Marshall AG. Reading chemical fine print: Resolution and identification of 3000 nitrogen-containing aromatic compounds from a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of heavy petroleum crude oil. Energy Fuels 2001;15:492–8.

[41] Burnham JF. Scopus database: a review. Biomed Digit Libr 2006;3:8.

[42] Rose ME, Kitchin JR. Pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. SoftwareX 2019;10:100263.

[43] AlRyalat SAS, Malkawi LW, Momani SM. Comparing bibliometric analysis using PubMed, Scopus, and Web of Science Databases. J Vis Exp 2019;152:12.

[44] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[45] Bishop CM. Pattern recognition and machine learning. Information science and statistics, 1st ed.. New York, NY: Springer; 2006.

[46] Mitchell TM. Machine learning. McGraw-Hill series in computer science. Artificial intelligence, 1st ed.. McGraw-Hill Education; 1997.

[47] Pearson K. On lines and planes of closest fit to systems of points in space. Phil Mag 1901;2:559–72.

[48] Fossen M, Hemmingsen PV, Hannisdal A, Sjöblom J, Kallevik H. Solubility parameters based on IR and NIR spectra: I. Correlation to polar solutes and binary systems. J Dispers Sci Technol 2004;26:227–41.

[49] Hur M, Yeo I, Park E, Kim YH, Yoo J, Kim E, No M-h, Koh J, Kim S. Combination of statistical methods and Fourier transform ion cyclotron resonance mass spectrometry for more comprehensive, molecular-level interpretations of petroleum samples. Anal Chem 2010;82:211–8.

[50] Chiaberge S, Fiorani T, Savoini A, Bionda A, Ramello S, Pastori M, Cesti P. Classification of crude oil samples through statistical analysis of APPI FTICR mass spectra. Fuel Process Technol 2013;106:181–5.

[51] Vaz BG, Abdelnur PV, Rocha WFC, Gomes AO, Pereira RCL. Predictive petroleomics: Measurement of the total acid number by electrospray Fourier transform mass spectrometry and chemometric analysis. Energy Fuels 2013;27:1873–80.

[52] Sad CM, d. Silva M, d. Santos FD, Pereira LB, Corona RR, Silva SR, Portela NA, Castro EV, Filgueiras PR, Jr VL. Multivariate data analysis applied in the evaluation of crude oil blends. Fuel 2018;239:421–8.

[53] Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In: Matrix pencils. Lecture notes in mathematics, vol. 973, Berlin, Heidelberg: Springer; 1983, p. 286–93.

[54] Terra LA, Filgueiras PR, Tose LV, Romão W, d. Souza DD, d. Castro EVR, d. Oliveira MSL, Diase JCM, Poppi RJ. Petroleomics by electrospray ionization FT-ICR mass spectrometry coupled to partial least squares with variable selection methods: prediction of the total acid number of crude oils. Analyst 2014;139:4908–16.

[55] Terra LA, Filgueiras PR, Tose LV, Romão W, d. Castro EV, d. Oliveira LM, Dias JC, Vaz BG, Poppi RJ. Laser desorption ionization FT-ICR mass spectrometry and CARSPLS for predicting basic nitrogen and aromatics contents in crude oils. Fuel 2015;160:274–81.

[56] Lozano DCP, Orrego-Ruiz JA, Hernández RC, Guerrero JE, Mejía-Ospino E. APPI(+)-FTICR mass spectrometry coupled to partial least squares with genetic algorithm variable selection for prediction of API gravity and CCR of crude oil and vacuum residues. Fuel 2017;193:39–44.

[57] Chua CC, Brunswick P, Kwok H, Yan J, Cuthbertson D, Aggelen Gv, Helbing CC, Shang D. Enhanced analysis of weathered crude oils by gas chromatography-flame ionization detection, gas chromatography-mass spectrometry diagnostic ratios, and multivariate statistics. J Chromatogr A 2020;1634:461689.

[58] Melendez-Perez JJ, Oliveira LFC, Miranda N, Sussulini A, Eberlin MN, Bastos WL, Rangel MD, d. S. Rocha Y. Lacustrine versus marine oils: Fast and accurate molecular discrimination via electrospray Fourier transform ion cyclotron resonance mass spectrometry and multivariate statistics. Energy Fuels 2020;8:9222–30.

[59] Tøndel K, Indahl UG, Gjuvsland AB, Vik JO, Hunter P, Omholt SW, Martens H. Hierarchical cluster-based partial least squares regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models. BMC Syst Biol 2011;5:90.

[60] Bishop CM. Neural networks for pattern recognition. Advanced texts in econometrics, 1st ed.. vol. 198, Madison Ave. New York, NY, United States: Oxford University Press, Inc.; 1995.

[61] Udelhoven T, Naumann D, Schmitt J. Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria. Appl Spectrosc 2000;54.

[62] Udelhoven T, Novozhilovx M, Schmitt J. The NeuroDeveloper®: a tool for modular neural classification of spectroscopic data. Chemometr Intell Lab Syst 2003;66:219–26.

[63] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

[64] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85–117.

[65] Elgibaly A, Elkamel A. A new correlation for predicting hydrate formation conditions forvarious gas mixtures and inhibitors. Fluid Phase Equilib 1998;152:23–42.

[66] Elgibaly A, Elkamel A. Optimal hydrate inhibition policies with the aid of neural networks. Energy Fuels 1998;13:105–13.

[67] Chapoy A, Mohammadi A, Richon D. Predicting the hydrate stability zones of natural gases using artificial neural networks. Oil Gas Sci Technol 2007;62:701–6.

[68] Ghavipour M, Ghavipour M, Chitsazan M, Najibi SH, Ghidary SS. Experimental study of natural gas hydrates and a novel use of neural network to predict hydrate formation conditions. Chem Eng Res Des 2012;91:264–73.

[69] Hesami SM, Dehghani M, Kamali Z, Bakyani AE. Developing a simple-to-use predictive model for prediction of hydrate formation temperature. Int J Ambient Energy 2015;38:380–8.

[70] Soroush E, Mesbah M, Shokrollahi A, Rozyn J, Lee M, Kashiwao T, Bahadori A. Evolving a robust modeling tool for prediction of natural gas hydrateformation conditions. J Unconv Oil Gas Resour 2015;12:45–55.

[71] Ghayyem MA, Nasab AG, Khormizi MZ, Rostami M. Predicting the conditions for gas hydrate formation. Pet Sci Technol 2019;37:1855–60.

[72] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20. 273-229.

[73] Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2:121–67.

[74] Cao J, Zhu S, Li C, Han B. Integrating support vector regression with genetic algorithm for hydrate formation condition prediction. Processes 2020;8:519.

[75] Qin H, Srivastava V, Wang H, Zerpa LE, Koh CA. Machine learning models to predict gas hydrate plugging risks using flowloop and field data. In: Offshore technology conference, conference paper. 2019, p. 12.

[76] Rashid S, Fayazi A, Harimi B, Hamidpour E, Younesi S. Evolving a robust approach for accurate prediction of methane hydrate formation temperature in the presence of salt inhibitor. J Nat Gas Sci Eng 2014;18:194–204.

[77] Mesbah M, Soroush E, Rezakazemi M. Development of a least squares support vector machine model for prediction of natural gas hydrate formation temperature. Chin J Chem Eng 2016;25:1238–48.

[78] Ghiasi MM, Yarveicy H, Arabloo M, Mohammadi AH, Behbahani RM. Modeling of stability conditions of natural gas clathrate hydrates using least squares support vector machine approach. J Mol Liq 2016;223.

[79] Yarveicy H, Ghiasi MM. Modeling of gas hydrate phase equilibria: Extremely randomized trees and LSSVM approaches. J Mol Liq 2017;243:533–41.

[80] Suykens J, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9:293–300.

[81] Quinlan JR. Simplifying decision trees. Int J Man-Mach Stud 1987;27:221–34.

[82] Utgoff PE. Incremental induction of decision trees. Mach Learn 1989;4:161–86.

[83] Brodley CE, Utgoff PE. Multivariate decision trees. Mach Learn 1995;19:45–77.

[84] Breiman L. Arcing classifier (with discussion and a rejoinder by the author). Ann Statist 1998;26(3):801–49.

[85] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40.

[86] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[87] Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 1998;20:832–44.

[88] Song Y, Zhou H, Wang P, Yang M. Prediction of clathrate hydrate phase equilibria using gradient boosted regression trees and deep neural networks. J Chem Thermodyn 2019;135:86–96.

[89] Acharya PV, Bahadur V. Thermodynamic features-driven machine learning-based predictions of clathrate hydrate equilibria in the presence of electrolytes. Fluid Phase Equilib 2021;530:112894.

[90] Lovatti BP, Nascimento MH, Rainha KP, Oliveira EC, Neto AC, Castro EV, Filgueiras PR. Different strategies for the use of random forest in NMR spectra. J Chemometr 2020;34:e3231.

[91] Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artif Intell Rev 2007;26:159–90.

[92] a. Ratanamahatana C, Gunopulos D. Feature selection for the naive bayesian classifier using decision trees. Appl Artif Intell 2003;17(5–6):475–87.

[93] Shi J, Li J, Usmani AS, Zhu Y, Chen G, Yang D. Probabilistic real-time deep-water natural gas hydrate dispersion modeling by using a novel hybrid deep learning approach. Energy 2021;219:119572.

[94] Zhang P, Chen X, Fan C. Research on a safety assessment method for leakage in a heavy oil gathering pipeline. Energies 2020;13:1340.

[95] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Amer Statist 1992;46:175–85.

[96] Xu H, Jiao Z, Zhang Z, Huffman M, Wang Q. Prediction of methane hydrate formation conditions in salt water using machine learning algorithms. Comput Chem Eng 2021;151:107358.

[97] Amin JS, Bahadori A, Nia BH, Rafiee S, Kheilnezhad N. Prediction of hydrate equilibrium conditions using k-nearest neighbor algorithm to CO2 capture. Pet Sci Technol 2017;35:1070–7.

[98] Hoerl AE. Application of ridge analysis to regression problems. Chem Eng Prog 1958;58(3):54–9.

[99] Tibshirani R. Regression Shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol 1996;58(1):267–88.

[100] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005;67:301–20.

[101] Cao K-AL, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol 2008;7:35.

[102] Sæbø S, Almøy T, Aarøe J, Aastveit AH. ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. J Chemometr 2007;22:54–62.

[103] Liland KH, Indahl U. Powered partial least squares discriminant analysis. J Chemometr 2009;23:7–18.

[104] Landgrebe MKB, Nkazi D. Toward a robust, universal predictor of gas hydrate equilibria by means of a deep learning regression. ACS Omega 2019;4:22399–417.

[105] Singh H, Seol Y, Myshakin EM. Prediction of gas hydrate saturation using machine learning and optimal set of well-logs. Comput Geosci 2020.

[106] Fu G-H, Zong M-J, Wang F-H, Yi L-Z. A comparison of sparse partial least squares and elastic net in wavelength selection on NIR spectroscopy data. Int J Anal Chem 2019;2019:7314916.

[107] Lv J, Wei J, Wang Z, Cao J. Multiple compounds recognition from the tandem mass spectral data using convolutional neural network. Molecules 2019;24:4590.

[108] Skarysz A, Alkhalifah Y, Darnley K, Eddleston M, Hu Y, Mclaren DB, Nailon WH, Salman D, Sykora M, Thomas CLP, Soltoggio A. Convolutional neural networks for automated targeted analysis of raw gas chromatography-mass spectrometry data. In: International joint conference on neural networks (IJCNN 2018). Rio de Janeiro, Brazil; 2018, p. 1–8.

[109] Kim S, Lee K, Lee M, Ahn T, Lee J, Suk H, Ning F. Saturation modeling of gas hydrate using machine learning with X-ray CT images. Energies 2020;13:5032.

[110] Li H, Misra S, He J. Neural network modeling of in situ fluid-filled pore size distributions in subsurface shale reservoirs under data constraints. Neural Comput Appl 2020;32:3873–85.