Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A liquid loading prediction method of gas pipeline based on machine learning

Bing-Yuan Hong [a, b], Sheng-Nan Liu [b, c], Xiao-Ping Li [b, **], Di Fan [d], Shuai-Peng Ji [b], Si-Hang Chen [b], Cui-Cui Li [a], Jing Gong [b, *]

[a] National-Local Joint Engineering Laboratory of Harbor Oil & Gas Storage and Transportation Technology/Zhejiang Provincial Key Laboratory of Petrochemical Pollution Control, School of Petrochemical Engineering and Environment, Zhejiang Ocean University, Zhoushan, 316022, China
[b] National Engineering Laboratory for Pipeline Safety/MOE Key Laboratory of Petroleum Engineering/Beijing Key Laboratory of Urban Oil and Gas Distribution Technology, China University of Petroleum-Beijing, Beijing, 102249, China
[c] China Huadian Group Energy Co., Ltd. North China Branch, Tianjin, 300280, China
[d] China Petroleum Engineering & Construction Corp, Beijing, 100120, China

ARTICLE INFO

Article history:
Received 26 August 2021
Received in revised form
5 May 2022
Accepted 5 May 2022
Available online 7 May 2022

Edited by Xiu-Qiu Peng

Keywords:
Liquid loading
Data-driven method
Machine learning
Gas pipeline
Multiphase flow

ABSTRACT

The liquid loading is one of the most frequently encountered phenomena in the transportation of gas pipeline, reducing the transmission efficiency and threatening the flow assurance. However, most of the traditional mechanism models are semi-empirical models, and have to be resolved under different working conditions with complex calculation process. The development of big data technology and artificial intelligence provides the possibility to establish data-driven models. This paper aims to establish a liquid loading prediction model for natural gas pipeline with high generalization ability based on machine learning. First, according to the characteristics of actual gas pipeline, a variety of reasonable combinations of working conditions such as different gas velocity, pipe diameters, water contents and outlet pressures were set, and multiple undulating pipeline topography with different elevation differences was established. Then a large number of simulations were performed by simulator OLGA to obtain the data required for machine learning. After data preprocessing, six supervised learning algorithms, including support vector machine (SVM), decision tree (DT), random forest (RF), artificial neural network (ANN), plain Bayesian classification (NBC), and K nearest neighbor algorithm (KNN), were compared to evaluate the performance of liquid loading prediction. Finally, the RF and KNN with better performance were selected for parameter tuning and then used to the actual pipeline for liquid loading location prediction. Compared with OLGA simulation, the established data-driven model not only improves calculation efficiency and reduces workload, but also can provide technical support for gas pipeline flow assurance.

1. Introduction

Natural gas contains a small amount of saturated water vapor changing with temperature and pressure during transportation in pipeline (Fan et al., 2021; Hong et al., 2020b). As the temperature drops, water vapor will condense out to form liquid water. Due to the large terrain undulation in mountainous areas, it's easy for liquid water to gather in some sections of the pipeline network (Hong et al., 2019), which will not only increase the friction in the pipeline, but also hinder gas flow, reduce gas transmission efficiency, and affect the economy of the gathering pipeline network (Shi et al., 2021). In addition, the gas-liquid two-phase flow is also prone to form slug flow, which can trigger fluctuations of pressure and flow in the pipeline and damage the pipeline (He et al., 2019). Therefore, it's essential to study the law of liquid loading in undulate pipelines and predict the situation of liquid loading for reducing production costs and ensuring the safety of transportation.

The root of liquid loading in the pipeline is gas-liquid two-phase

* Corresponding author.
** Corresponding author.
E-mail addresses: hongby@zjou.edu.cn (B.-Y. Hong), xpli126@126.com (X.-P. Li), ydgj@cup.edu.cn (J. Gong).

https://doi.org/10.1016/j.petsci.2022.05.002

flow. In the two-phase flow, there are both external forces between the fluid and the inner wall of the pipe, as well as interaction forces between the two-phase interface (Chen et al., 2021a; Shi et al., 2021). The mechanical relationship of external forces and interaction forces can be affected by the flow pattern (Izwan Ismail et al., 2015). The flow pattern of multiphase flow is changeable, which is closely related to problems such as the energy loss of the fluid in the tube (He et al., 2018). The early scholars mainly studied the flow patterns experimentally and tried to propose generalized flow pattern diagrams (Khaledi et al., 2014; Ong and Thome, 2011). As the research progresses, some scholars began to use mathematical models to explore the relationship between various multiphase flow parameters and flow pattern changes, and established different flow patterns judgment criteria (Barnea, 1987; Taitel and Dukler, 1976). The flow patterns discrimination is the prerequisite for calculating the liquid holdup and pressure drop. For the study of liquid holdup, many scholars proposed models and related formulas based on empirical and semi-empirical formulas (Zhang et al., 2004). For the study of pressure drop, many scholars derived pressure drop calculation formulas applicable to different pipeline structures based on energy conservation by considering pipeline undulation and downhill pipeline energy recovery (Rodrigues et al., 2018). In addition, some scholars employed numerical simulation methods to explore the law of liquid loading (Ming et al., 2018; Vieira et al., 2021), and analyzed in detail the factors that affect the generation of liquid loading during transportation (Abubakar et al., 2018; Kesana et al., 2018; Rodrigues et al., 2020).

The above researches revealed the law of liquid loading from different aspects, analyzed the changes of pressure drop, liquid holdup and other indicators with different parameters such as flow velocity, pipe diameter, water content, and outlet pressure, and established prediction models such as the critical inclination angle of the pipeline (Liang et al., 2021; Salubi et al., 2021). However, most of the single models do not consider the cross-effects of multiple factors, so the application range is limited, the conclusions obtained are not strong in regularity, and the accuracy of pipeline liquid load prediction is still not good enough. The complexity of the physical flow process makes it particularly difficult to establish the mathematical model for the liquid loading process. Unlike a single model, the commercial simulator OLGA combines a variety of relational expressions to realize the coverage of multiple working conditions, and the results obtained have been tested in practice that can be used in engineering applications (Shi et al., 2020). Nevertheless, OLGA only simulates specific working conditions and requires separate recalculations for different working conditions each time (Kanin et al., 2019). In fact, the operating parameters and pipeline parameters are complex and diverse in engineering practice. If numerical simulation modeling such as OLGA is required every time, the workload is large and the efficiency is low. Therefore, it is necessary to develop a new method that can quickly predict liquid loading.

The development of big data technology and artificial intelligence provides the possibility to establish data-driven models. Machine learning (ML) methods are widely used in various complex engineering problems in many fields. Qi et al. (2018) used artificial neural network (ANN) and particle swarm algorithm to predict the unconfined compressive strength of cemented paste backfill. The data was obtained through experiments, and the minimum mean square error (MSE) and correlation coefficient (R) were used to evaluate the performance of the optimal ANN model on the training set and the test set. Kanin et al. (2019) proposed a ML algorithm for steady-state simulation of multiphase pipelines based on laboratory data. Three models were trained using various ML algorithms on representative laboratory datasets selected from the open literature. The first model was used to predict the liquid holdup, the second model was used to determine the flow pattern, and the third model was used to estimate the pressure gradient. It has been verified that the models can be extended from the laboratory to field conditions. Mask et al. (2019) established a new model with the help of ML by dimensional analysis of more than 8000 laboratory multiphase flow tests. The test results show that the flow pattern is affected by the fluid properties, the field flow rate of fluids, the geometry and mechanical properties of the flow conduit. Moreover, ML technology has significantly improved the prediction accuracy of dimensionless variables compared with semi-analytical models. Lin et al. (2020) reported a method for predicting the flow patterns of upwardly inclined pipes through deep learning neural networks. The data came from experimental data sets reported in the literature, and the surface velocity and inclination of single phase were selected as input parameters to identify the flow pattern. Compared with the classic flow pattern diagram, the effectiveness of the prediction model was verified. The above researches not only have promoted the continuous development and practical application of ML but also cover various fields and provide a new idea and method for the study of liquid loading.

However, to the best of our knowledge, there are few studies on the use of ML in liquid loading prediction. Therefore, on the basis of numerical simulation by using the simulator, this paper proposed a ML method to establish a natural gas pipeline liquid loading prediction model with high generalization ability to predict pipeline liquid loading. First, according to the characteristics of actual gas pipeline, a variety of reasonable combinations of working conditions such as different gas velocities, pipe diameters, water contents and outlet pressures were set, and multiple undulating pipeline topography with different elevation differences was established. Then a large number of simulations were performed by simulator OLGA to obtain the data required for machine learning. After data preprocessing, six supervised learning algorithms, including support vector machine (SVM), decision tree (DT), random forest (RF), artificial neural network (ANN), plain Bayesian classification (NBC), and K nearest neighbor algorithm (KNN), were compared to evaluate the performance of liquid loading prediction. Finally, the RF and KNN with better performance were selected to optimize the parameters and applied to the actual pipeline for liquid loading location prediction.

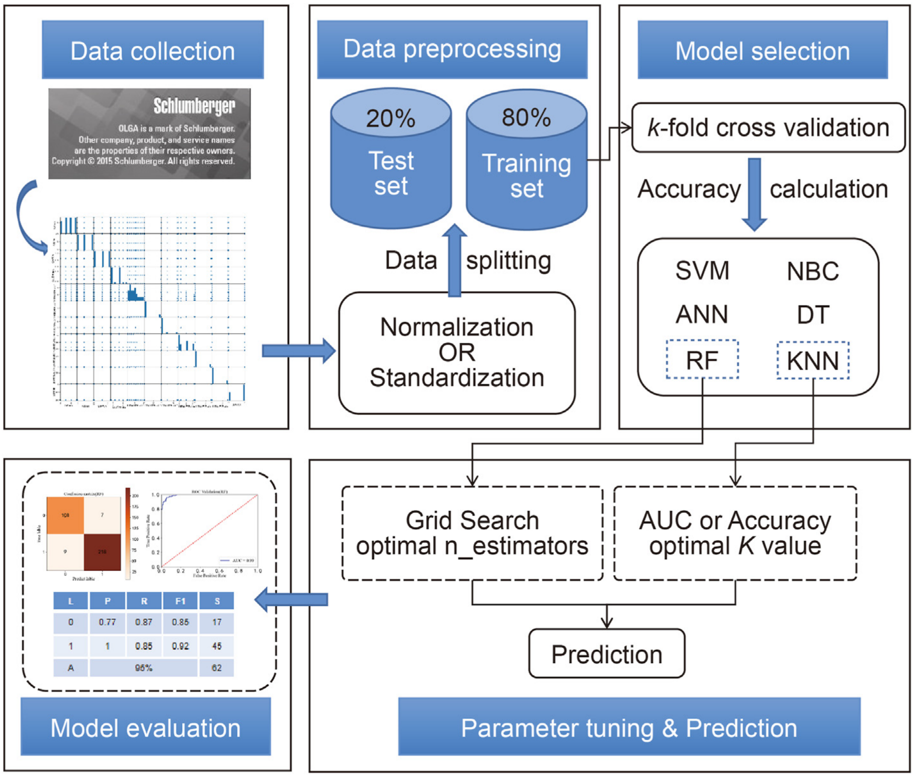
The paper is organized as follows: Section 2 describes the proposed liquid loading prediction method of gas pipeline based on machine learning, including data collection, data preprocessing, machine learning algorithms and model performance evaluation. The performance of six different Machine Learning algorithms is compared and analyzed in Section 3. Based on the performance, RF and KNN are selected to determine the liquid loading of gas pipeline. Finally, conclusions are drawn in Section 4.

2. Methodology

The proposed methodology based on ML follows the following five steps: data collection, data preprocessing and partitioning, model selection, parameter tuning and prediction, model evaluation and visual analysis, as shown in Fig. 1.

2.1. Data collection

In the research of pipeline liquid loading, the description of topography has always been a difficult problem. The inclination, length and elevation of pipeline laying can affect the generation of liquid loading (He et al., 2018; Ming et al., 2018). As shown in Fig. 2, for the upward dipping section of the pipeline, the liquid layer has a tendency to move to the lower part of the pipeline due to the



L	P	R	F1	S
0	0.77	0.87	0.85	17
1	1	0.85	0.92	45
A	95%			62

Fig. 1. Flow chart of prediction method for liquid loading based on machine learning.

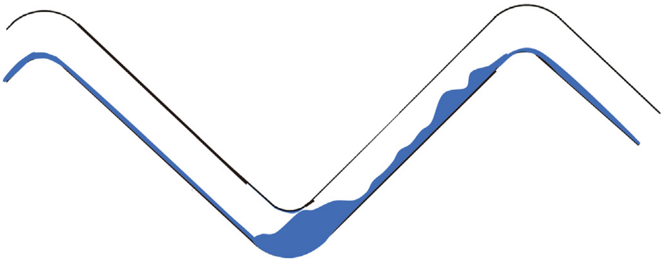


Fig. 2. Schematic diagram of fluid loading in natural gas undulating pipeline.

gravity after the liquid is completely flat on the upward dipping section of the pipeline; on the other hand, the pressure drop increases, as the gas expands, the flow rate increases, so the slip ratio between the gas and liquid phases increases, which will reduce the liquid carrying capacity of the gas phase, resulting in a tendency for the liquid phase to move to the low part, thereby forming an accumulation of liquid at the low point of the pipeline. In the downward-dip section of the pipeline, the liquid phase flows downward under the gravity, while the flow rate of the gas phase decreases under the action of buoyancy, which will reduce the slip ratio between the two phases and increase the liquid-carrying capacity of the gas phase. Therefore, there will not be fluid accumulation in the downdip section and the liquid loading phenomenon usually occurs in the climbing section of the pipeline, so the topography of the pipeline in this paper is characterized by the combination of the inclination angle and the mileage of the upward pipe.

In this paper, the data needed for ML is obtained by OLGA simulator which has been validated by many scholars and has been widely used in the numerical simulation of multiphase flow. 270 groups of three-stage undulating pipeline examples, as shown in Fig. 3-a and Fig. 3-b, were designed by OLGA simulation as a part of the data set. Meanwhile, because liquid loading usually occurs in the upward pipe, in order to predict the situation of liquid loading in the condition of different pipeline inclination angles, 1440 sets of single undulating pipeline terrains with various upward pipe inclination α and mileage L were created by OLGA simulation, as shown in Fig. 3-c. It can be considered that the inclination angle α and mileage L of the second section of the pipeline and above are 0. The pipeline material is PE, the wall roughness is 10 μm, the total heat transfer coefficient is 1.75 $\text{W/m}^2\cdot°\text{C}$, and the pipeline inlet temperature is 10 °C. The specific parameter settings are shown in Tables 1–3. It should be noted that the gas pressure range in Table 1 and the components in Table 3 are consistent with the field. Most of the flow patterns in the ML data set are stratified flow. However, the flow pattern, like the liquid loading, is the output result under the complex relationship of many factors, such as pipe diameter, gas-liquid velocity, pipe inclination and so on. In addition, there are many different ways to define the flow pattern, and the same phenomenon can get different flow patterns according to different division methods. Therefore, the flow pattern is not used as an input parameter of the model. With the same mass flow rate but different pipe diameters, the gas velocity will be different, and the liquid-carrying capacity is definitely different. Different liquid carrying capacity will affect the results of liquid loading, so gas velocity is chosen instead of mass flow rate.

Based on the OLGA simulation results, “whether it accumulates liquid” was taken as the target value. If liquid loading occurs, it is mapped to the value of “1”, otherwise “0”. A CSV file was created and read into Python to obtain the characteristic attributes including setting parameters and target values. The distribution of each characteristic attribute can be obtained from matrix scatter diagram shown in Fig. 4. The diagonal line shows histograms, and

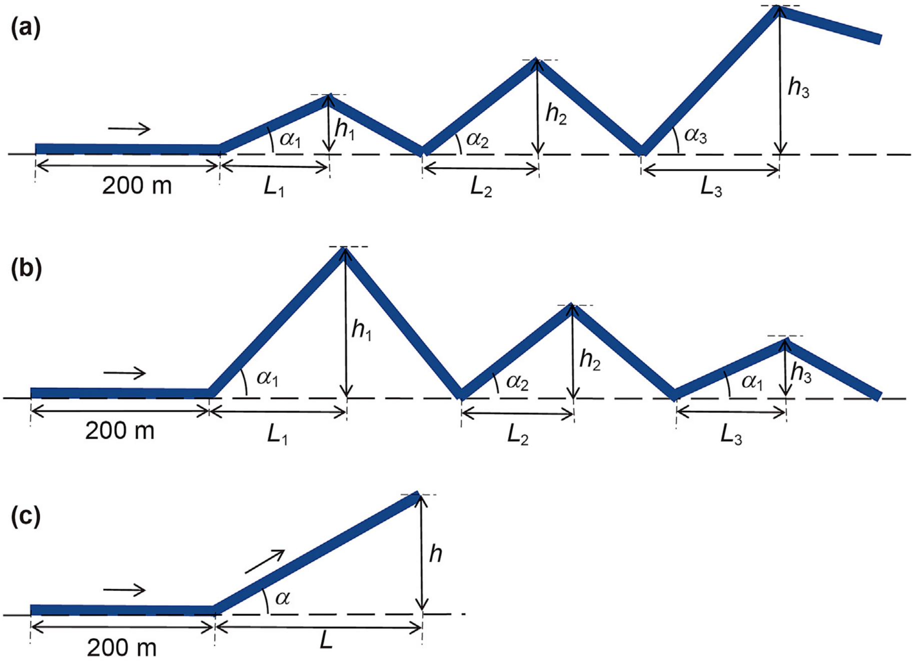


Fig. 3. Schematic diagram of pipeline, (**a**) three-stage undulating with terrain rising, (**b**) three-stage undulating with terrain decreasing, (**c**) single-stage undulating.

Table 1
Parameter setting of liquid loading numerical simulation for data collection.

Parameter	Setting range
Outlet pressure (p), MPa	0.04, 0.06, 0.08, 0.1, 0.12
Gas saturated water content (w)	0.5%, 0.8%, 1.1%
Gas velocity (V), m/s	10, 15, 20, 25
Pipe diameter (d)	DN100, DN150, DN200
Inclination angle (α)	5–50°, take a value every 5°
Mileage of upward inclined pipe (L), m	200, 500, 800
Elevation (h)	Calculated according to inclination angle and updip pipe mileage, as shown in Table 2

Table 2
Pipeline elevation data.

Inclination, °	Mileage, m	Elevation, m	Inclination, °	Mileage, m	Elevation, m
5	200	17.50	20	500	181.99
5	500	43.74	20	800	291.18
5	800	69.99	25	200	93.26
10	200	35.27	25	500	233.15
10	500	88.16	30	200	115.47
10	800	141.06	30	500	288.68
15	200	53.59	35	200	140.04
15	500	133.97	40	200	167.82
15	800	214.36	45	200	200.00
20	200	72.79	50	200	238.35

Table 3
Gas composition parameter.

Composition	CH_4	C_2H_6	CO_2	N_2	Total
Volume fraction, %	99.02	0.01	0.26	0.71	100

the off-diagonal line shows scatter plots. Specifically, the diagonal part represents the distribution of the i-th feature, the x-axis is the value of the feature, and the y-axis is the number of occurrences of the feature's value, thus representing the density estimate of the i-th feature. The distribution of gas flow velocity, pipe inner diameter, water content, outlet pressure, inclination angle of each section of the upward pipe and the corresponding mileage of the upward pipe are shown from top to bottom. The last one in the diagonal part is whether it accumulates liquid, where “0” means no liquid loading, and “1” means liquid loading. A total of 1710 groups of samples, of which the number of samples without liquid loading was 574, and the number of samples with liquid loading was 1136, with a ratio of about 1:2. The off-diagonal part of the i-th row and j-th column represents the scatter plot of the i-th feature and the j-th feature. For example, the second row and the first column represent the parameter combination of gas velocity and pipe diameter. The velocity includes 10 m/s, 15 m/s, 20 m/s, and 25 m/s, and the pipe

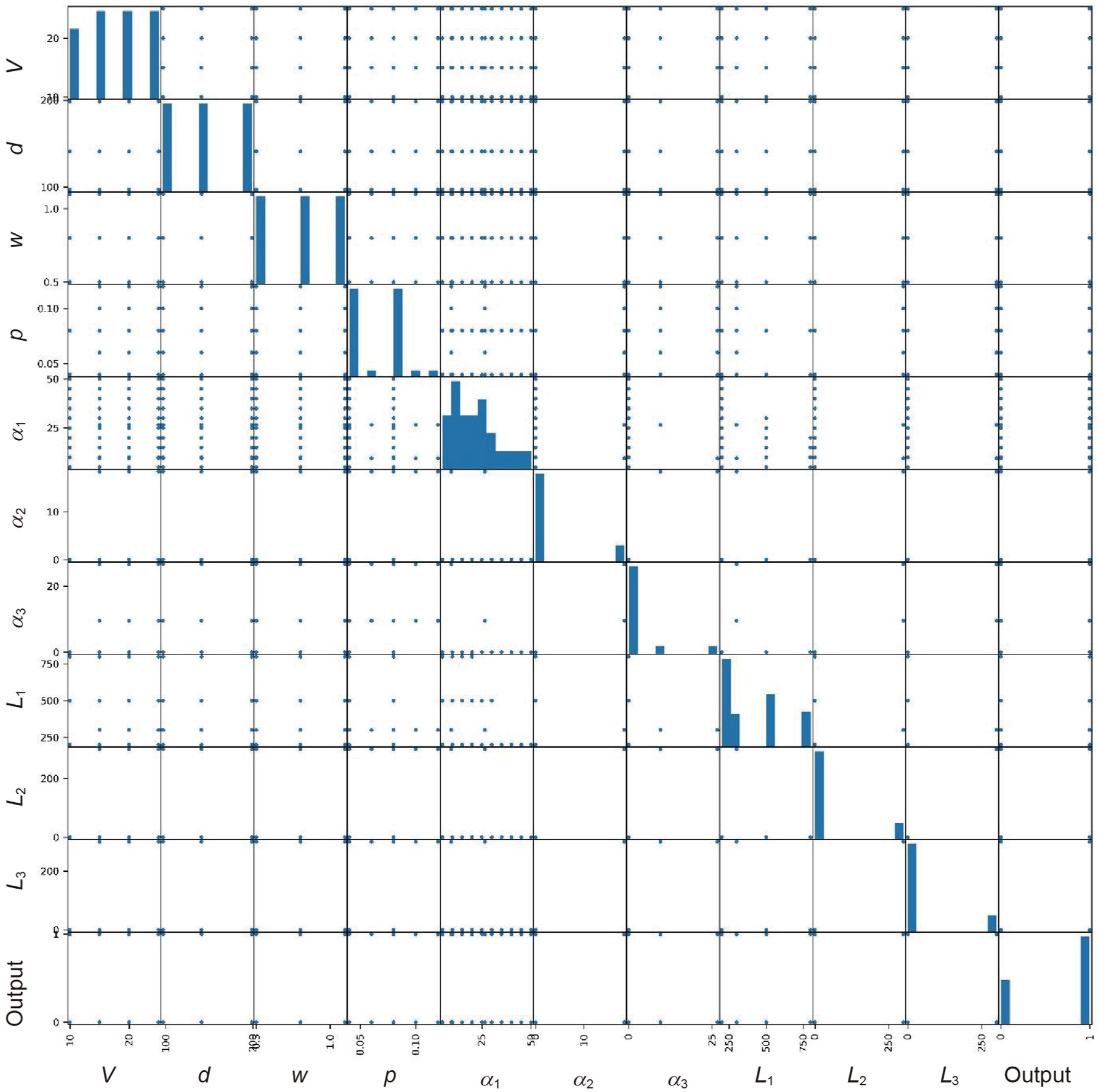


Fig. 4. Matrix scatter diagram of all attribute.

diameter includes DN100, DN150, and DN200. Since this data set is set according to certain rules, the distribution is relatively regular.

2.2. Data preprocessing

The data in this paper are produced from OLGA simulation and are clean. However, measurement noise is inevitable in practice, direct use of low-quality data for model training will lead to low-quality prediction results. Therefore, the influence of noise data should be identified and reduced through data cleaning. Since the dimensions and units of each attribute are different, it needs to be standardized. Z-Score standardization (Finch and Beck, 2011) is employed which converts the mean of the original data into 0 and the standard deviation into 1. The calculation formula is as shown in formula (1).

$$x^* = \frac{x - \overline{x}}{\sigma} \tag{1}$$

where x is the observed value, $\overline{x}$ is the overall average, and σ is the overall standard deviation.

The data was divided into a training set and a test set with a ratio of 8:2 using the hold-out. Set the parameter stratify = y, that is, allocate data according to the proportion of each category in the

Table 4
Data set division results.

Data set	Category	Quantity	Percentage of this category in the sample	Category 0: Category 1
Training set	0	459	79.97%	1:2
	1	909	80.02%	
Test set	0	115	20.03%	1:2
	1	227	19.98%	

original data label *y*, so that the proportion of each category data in the training set and the test set were the same as the original data set. The final division result is shown in Table 4 where category 0 means without liquid loading while category 1 means liquid loading.

2.3. Machine learning algorithms

The goal of this paper is to predict the liquid loading of gas pipeline by train the model through a large amount of known data. Therefore, the form of supervised learning was chosen. The concepts and principles of six commonly used supervised learning algorithms are shown in Table 5.

Random forest (RF) is a combination of Bootstrap aggregating algorithm and decision tree, which integrates multiple classifiers into a whole. The basic concept of a RF is to independently build several decision trees on random subsets of the original training dataset. The schematic diagram of RF is shown in Fig. 5, and it becomes a forest style vividly. Firstly, the bootstrap method is used to randomly select *k* rounds of samples from the data set with replacement to obtain k training sets; then, randomly extract a part of the feature attributes from each training set and apply them to node splitting, construct a decision tree, and finally construct *k* decisions tree; finally, the best classification is selected by "voting".

K-nearest neighbor algorithm is a typical supervised learning algorithm. In fact, the samples to be predicted are put into the data set, and *K* sample data closest to the sample to be predicted are obtained from the training set; the target attribute value of the current sample to be predicted is predicted according to the obtained *K* sample data. In a feature space, a sample also belongs to a class if most of its *K*-nearest neighbors belong to that class, as shown in Fig. 6. It can be used in both classification applications and regression applications. In classification prediction, the majority voting method is generally used; while in regression prediction, the average method is generally used. The algorithm is described as follows: 1) calculate the distance between the test data and each training data; 2) Sort according to the increasing relationship of distance; 3) Select the *k* points with the smallest distance; 4) Determine the occurrence frequency of the category where the first *K* points are located; 5) Return the category with the highest frequency in the first *K* points as the prediction classification of test data.

2.4. Model performance evaluation

The six algorithms in Table 5 are all available for classification, and the accuracy of different algorithms is evaluated by following formula on the same data set to select the most suitable algorithm.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where *TP* stands for true positive, *TN* stands for true negative, *FP* stands for false positive, *FN* stands for false negative.

To reduce the occasionality of calculation and the impact of data partition, the *k*-fold cross-validation method (Saud et al., 2020) was used to calculate the model accuracy. As shown in Fig. 7, the data set was divided equally into *k* pieces, and one of them was taken as the test set each time, the remaining was used as the training set.

Table 5
Classification of supervised learning algorithms.

Algorithms	Concept and principle	Advantage	Disadvantage	Application scenario
Support Vector Machines (SVM) (Lee, 2021; Zhang et al., 2021)	The principle of this method is to use the line or the surface as the decision boundary to classify the data binary.	Strong generalization ability	It is difficult to solve when the sample size is large; it is not suitable for multi-classification problems.	Facial recognition, text classification, biomedical diagnosis, etc.
Decision tree (DT) (W. Huo et al., 2021; Yuvaraj et al., 2021)	The feature attribute is used as the dividing node, and the sample is divided layer by layer from top to bottom, until the sample classification is obtained.	Can handle irrelevant features; easy to get visual results; easy to understand and analyze.	Prone to overfitting	Multiple classification problems
Random forest (RF) (Y. Huo et al., 2021; Tiwary et al., 2020)	Algorithm integrated by multiple decision trees.	Strong anti-jamming ability	Slower execution	Multiple classification and regression problems
Artificial neural networks (ANN) (Shi et al., 2021; Si et al., 2021)	A mathematical model based on biological neural networks.	Strong non-linear fitting ability; the rules are simple and easy to implement.	It is difficult to understand its internal operating mechanism.	Voice recognition, medical treatment, etc.
Naive Bayes Classification (NBC) (Andrejiova and Grincova, 2018; Khajenezhad et al., 2021)	Based on the knowledge of probability statistics, calculate the probability that the sample to be tested belongs to each category, and use the category with the highest probability as the category of this sample.	Simple logic; low false positive rate.	The prediction effect is poor for samples with high attribute relevance.	Text classification, face recognition, etc.
K nearest neighbor algorithm (KNN) (Hashemizadeh et al., 2021)	Set a reasonable K value, calculate the distance between the sample to be tested and the training sample, and get the K samples closest to it. The category with the highest frequency is the sample category to be tested.	The principle is simple; the accuracy is high.	The prediction effect is poor when the sample categories are not balanced.	Mail classification, image recognition, etc.

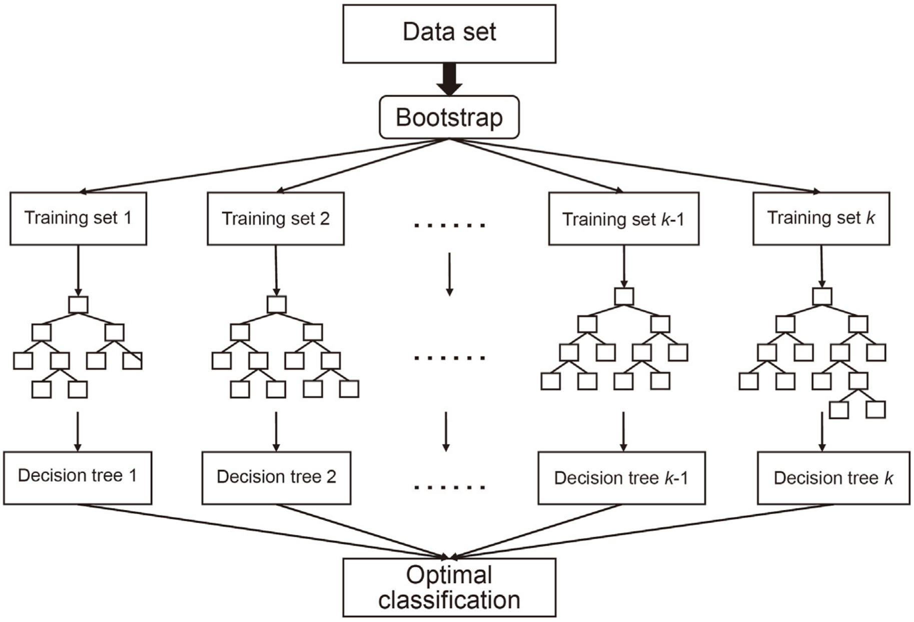


Fig. 5. The schematic diagram of random forest.

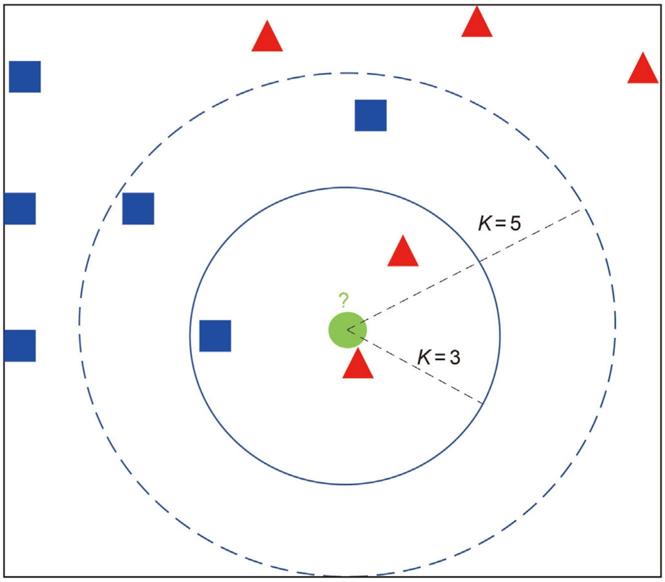


Fig. 6. The schematic diagram of *K*-nearest neighbor algorithm.

Therefore, a total of k times of training and testing was performed, and each time a result (usually the accuracy) that could evaluate the model performance was obtained. The average of the *k* times results was used as the final result. In this paper, $k = 5$ was taken to obtain the mean and standard deviation of accuracy of each algorithm, so as to conduct evaluation.

3. Result and discussion

3.1. Model performance comparison

The 10 variables shown in Fig. 4 were used as inputs to the above six static models, and the output was whether it accumulates liquid. Based on the data set and the partition method, the six supervised learning algorithms were model-trained. In order to reduce the impact of data division, the *k*-fold cross validation method is used for k times of training and testing. The accuracy of the model can be obtained by each time of calculation. The mean and standard deviation of the accuracy of k times are the final results. In this paper, $k = 5$ is used, and the mean and standard deviation of the accuracy of each algorithm are shown in Fig. 8. It can be seen from Fig. 8 that the accuracy of RF, DT, KNN and ANN

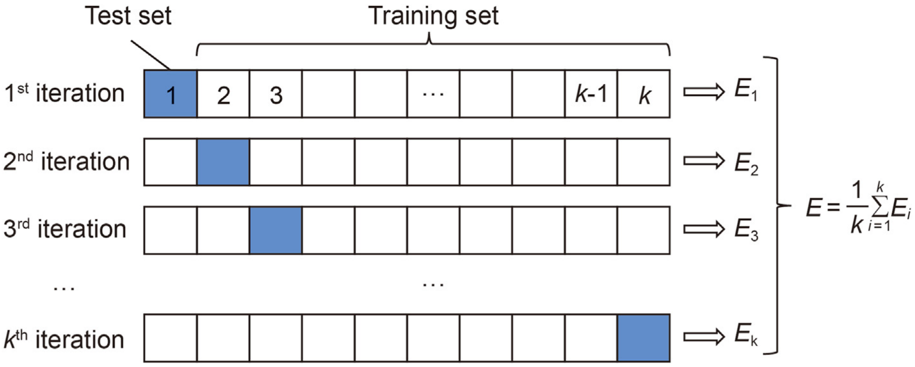


Fig. 7. Schematic diagram of *k*-fold cross-validation.

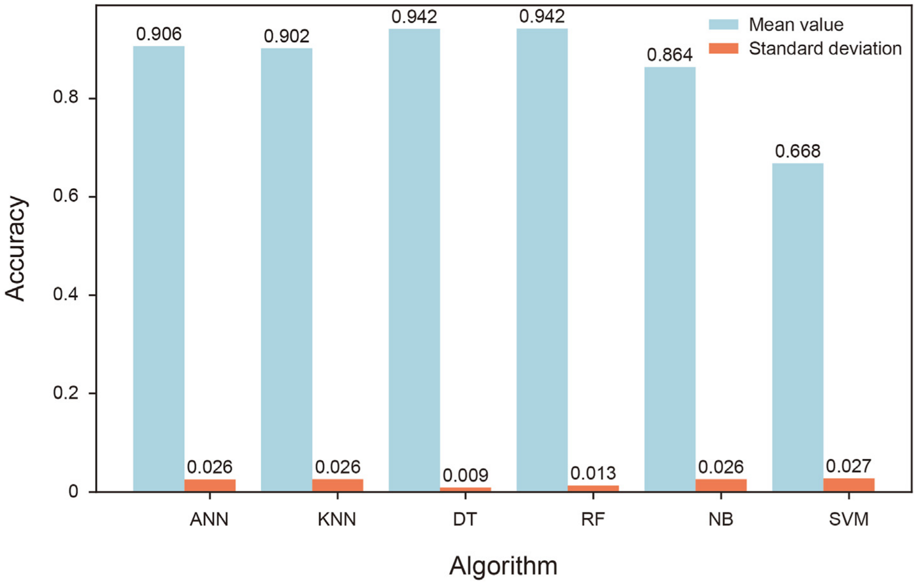


Fig. 8. Accuracy rate of each algorithm.

algorithms are all high. Among them, the RF model has the highest accuracy and has the following advantages over DT. (1) The influence of outliers on model prediction is reduced. According to the principle of RF, only part of the characteristic data is selected when generating each tree to finally build several different DTs and obtain multiple prediction results. Therefore, when individual outliers appear, the influence will not be great and the over-fitting probability of the model can be reduced at the same time. (2) The accuracy of the model is improved. Compared with the single DT, there are many "choices" in RF, and the best classification can be selected by comparing and analyzing the results of multiple DTs. Consequently, although DT is the second accurate algorithm, it can be replaced by RF. Moreover, from the perspective of learning methods, KNN algorithm is different from other supervised algorithms. It is a "lazy" learning algorithm, that is, it does not generate a classification or prediction model in advance for the prediction of new samples, but carries out the model construction and the prediction of unknown data at the same time. It has the advantages of simple principle and insensitivity to abnormal points. Therefore, based on the above analysis, the RF algorithm and KNN algorithm can be selected for parameter tuning and then used for pipeline liquid load prediction.

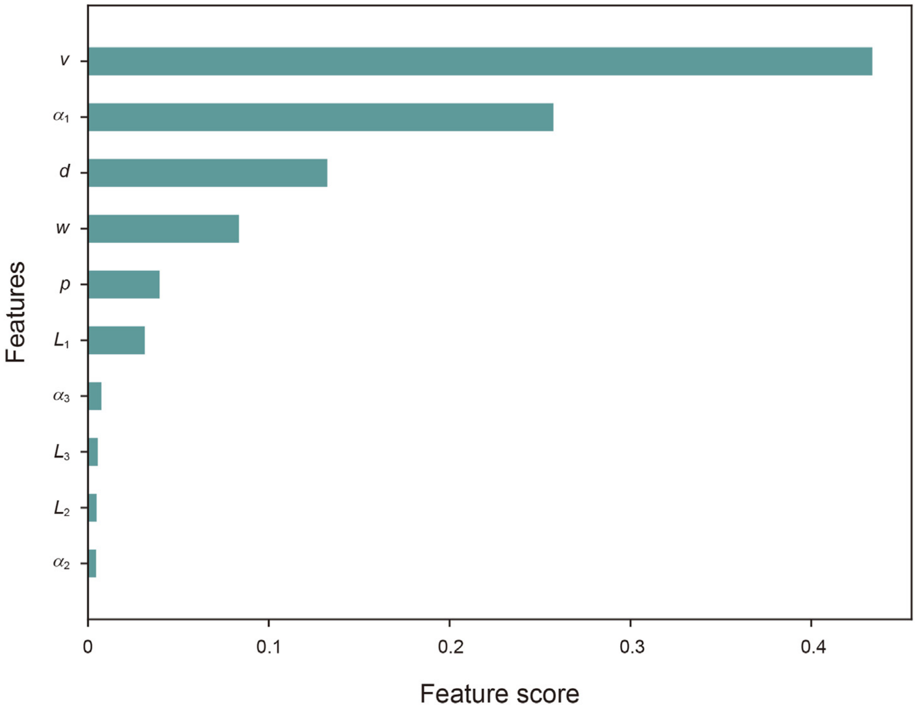


Fig. 9. Feature importance based on Gini coefficient.

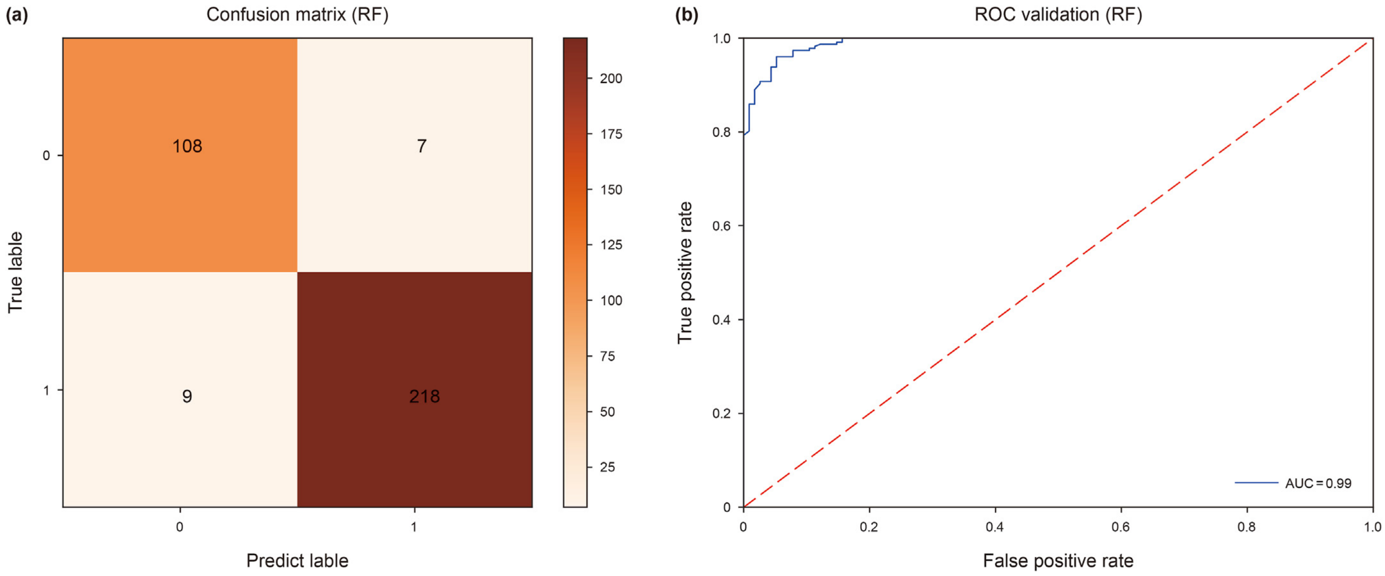


Fig. 10. Performance of RF algorithm, (**a**) Confusion matrix (**b**) ROC curve.

3.2. Model parameter tuning of RF and KNN

RF is the combination of Bootstrap Aggregating algorithm and DT, which integrates multiple classifiers into a whole (Hashemizadeh et al., 2021). The method of grid search combined with *k*-fold cross validation is adopted to determine the optimal number of classifiers (n_estimators). The optimal number of classifiers is obtained by setting the search range from 25 to 500 and the step size to 25, so n_estimators = 175 is used for training the model. Node splitting of DT in RF requires selecting a certain feature as the division attribute. The data set contains multiple features, and each feature has a different contribution to each tree. Hence, the feature used as the partition attribute usually is selected according to the reduction degree of Gini coefficient (Brown and Myles, 2020; Jain et al., 2021) before and after splitting. According to the results of Gini coefficient, the order of feature importance is: flow rate > inclination angle of first upper inclined pipe α_1>pipe inner diameter > water content > outlet pressure > mileage of first upper inclined pipe L_1>inclination angle of third upper inclined pipe α_3> mileage of third upper inclination pipe L_3> mileage of second upper inclination pipe L_2> inclination angle of second upper inclination pipe α_2, as is shown in Fig. 9. Therefore, follow the order of feature importance in Fig. 7 as the node splitting attribute of the DT.

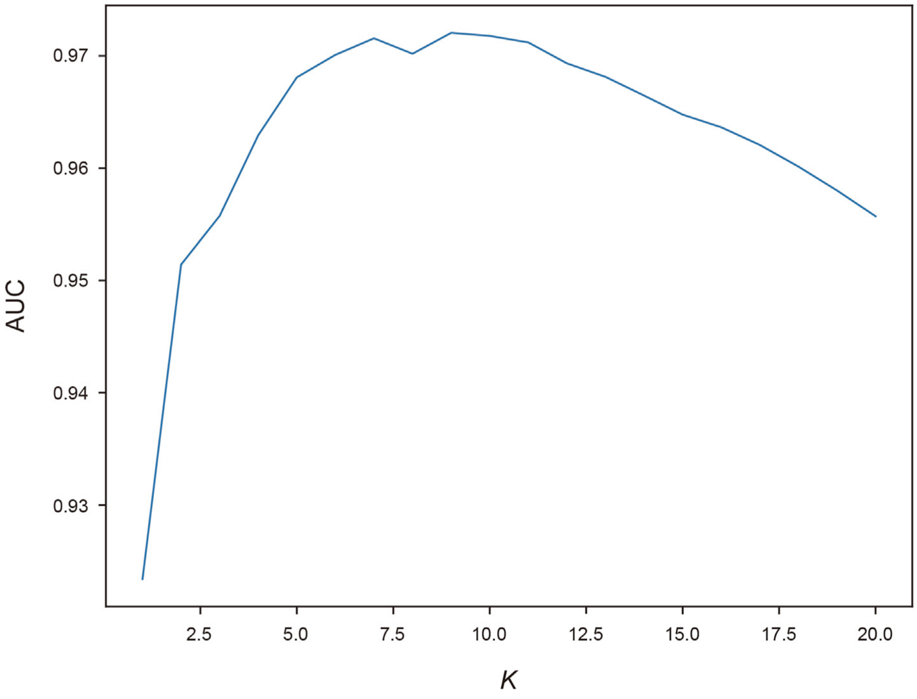


Fig. 11. AUC diagram of the KNN model under different *K* values.

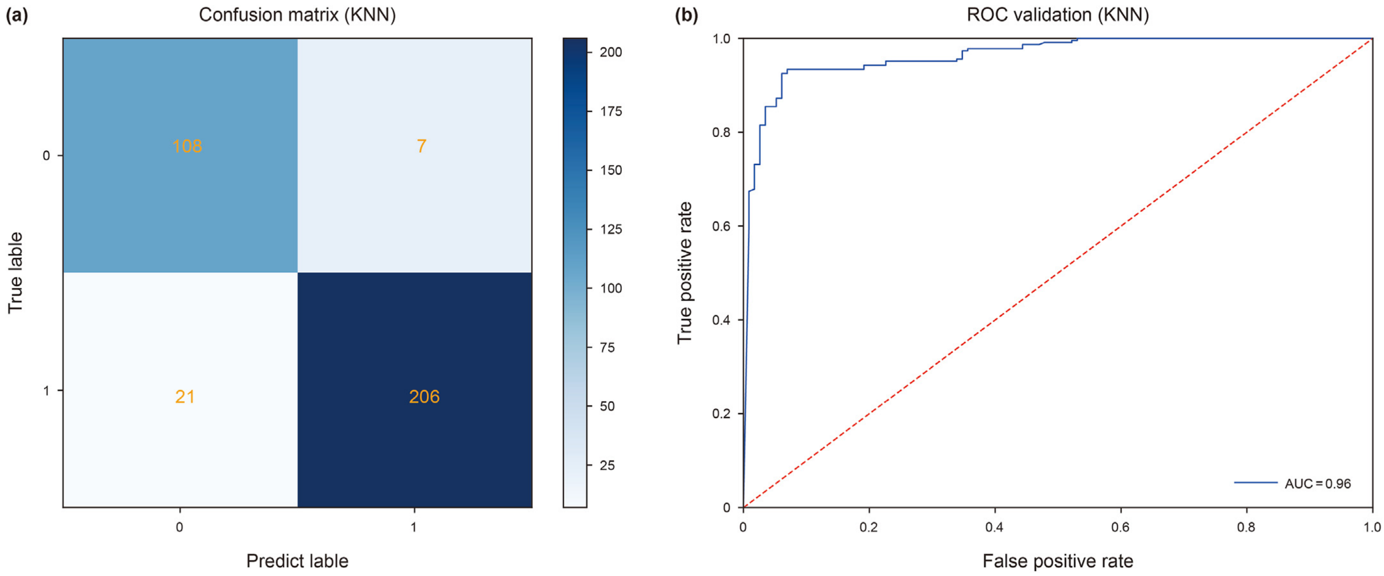


Fig. 12. Performance of KNN algorithm, (**a**) Confusion matrix (**b**) ROC curve.

Table 6
Numerical mapping of category labels.

First site of liquid loading	Mapped value
No fluid accumulation	0
The first low location	1
The second low location	2
The third low location	3
The first two low locations	4
The last two low locations	5

The trained RF model is applied to the test set, and the confusion matrix of the test results is shown in Fig. 10-a, and the ROC curve is shown in Fig. 10-b. The prediction is correct in 326 groups, including 108 groups with no liquid loading and 218 groups with liquid loading. There are 16 groups of incorrect predictions, among which 7 groups of no liquid loading samples are predicted as liquid loading and 9 groups of liquid loading samples are predicted as no liquid loading samples. The precision, recall rate, F1 value, and accuracy are all above 90%, and the AUC is 0.99. It can be seen that the RF model has strong ability to recognize samples and has a better prediction effect, so there is no need to tune other parameters. In addition, the out-of-bag score of the model is 0.9598, which indicates that the generalization ability of the model is better.

KNN is a classification that relies on distance calculation (Dong et al., 2021; Hashemizadeh et al., 2021). In this paper, the most widely used Euclidean distance is used to calculate the relationship between the sample predicted and the known sample. To improve the accuracy of the model, the cross-validation method is used to tune the parameter n_neighbors for determining the best K value, and the AUC values of the model under different K values are obtained as shown in Fig. 11 where the AUC value is the highest when the K value is 9. Therefore, $K = 9$ is selected for calculation.

The trained KNN model is applied to the test set, and the test results are shown in Fig. 12-a. The prediction is correct in 314 groups, including 108 groups with no liquid loading and 206 groups with liquid loading. There are 28 groups of incorrect predictions, among which 7 groups of no liquid loading samples are predicted as liquid loading and 21 groups of liquid loading samples are predicted as no liquid loading samples. The accuracy of the model is 92%, the precision, recall rate, and F1 value of the category “liquid loading” are all above 90%, and the accuracy of the category “no liquid loading” is slightly lower. The performance of the KNN model is slightly worse than that of the RF model, but the overall predictive ability is better. The ROC curve is shown in Fig. 12-b, indicating that the KNN model has a better prediction effect, with an AUC of 0.96, and there is no need for other parameter tuning.

3.3. Model application of RF and KNN

In the natural gas gathering pipelines, the installation of condensers is usually adopted to discharge the liquid loading (He et al., 2018). However, due to the complex and varied gas transportation conditions and pipeline terrains of the multi-undulating wet gas pipelines, the liquid loading is not always generated along the pipeline from front to back in order, but may be generated in the back section first (Chen et al., 2021b). With the increase of liquid loading in the back section of pipeline, the energy loss of the pipeline increases, resulting in liquid loading in the front section after a certain time. Therefore, if the location of the first liquid loading in the pipeline can be predicted and install a condenser here, the liquid loading in other locations of the pipeline could be avoided to a certain extent, and the liquid discharge could be

Table 7
Classified report of RF model prediction.

Category label	Precision	Recall	F1-score	Support
0	0.73	0.94	0.82	17
1	0.92	0.85	0.88	27
2	1.00	1.00	1.00	1
3	0.00	0.00	0.00	2
4	1.00	0.83	0.91	6
5	1.00	1.00	1.00	1
Accuracy	0.85			54

Table 8
Classified report of KNN model prediction.

Category label	Precision	Recall	F1-score	Support
0	0.77	0.87	0.85	17
1	1.00	0.85	0.92	27
2	1.00	1.00	1.00	1
3	1.00	0.50	0.67	2
4	1.00	1.00	1.00	6
5	1.00	1.00	1.00	1
Accuracy	0.91			54

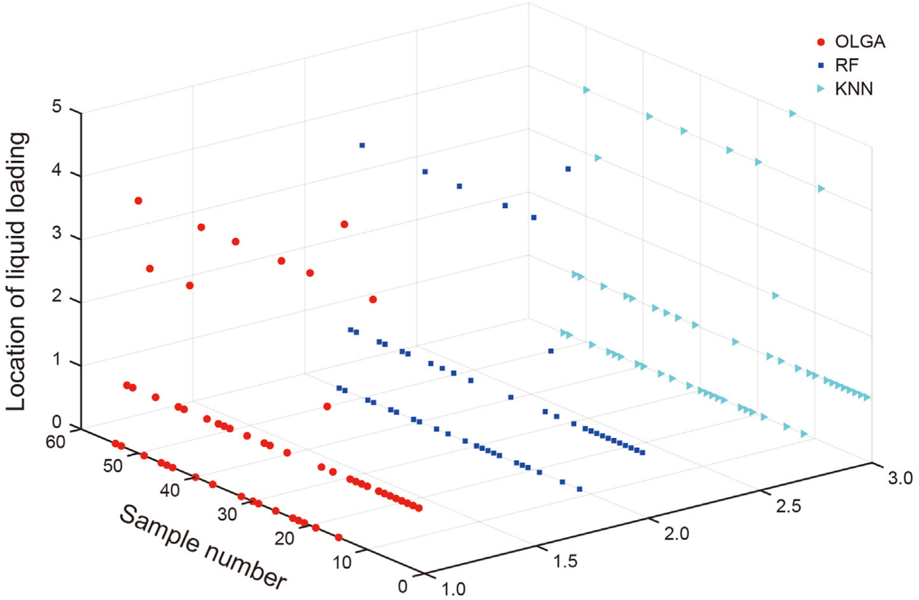


Fig. 13. Comparison diagram of machine learning model prediction and OLGA simulation.

maximized to improve the gas transmission efficiency of the pipeline (Hong et al., 2019, 2020a).

The above two algorithms RF and KNN are used to predict the location of the first liquid loading in the pipeline. Taking the “first liquid loading position” as the target value, six cases are analyzed and mapped into numerical form, as shown in Table 6.

Through the method of grid search combined with *k*-fold cross-validation, the optimal number of classifiers (n_estimators) is determined to be 450. There are 46 groups of correct predictions by the RF model and 8 groups of wrong predictions. The accuracy of the model and other indicators is shown in Table 7. It can be seen from the Table 7 that the accuracy of the RF model for predicting the location of liquid loading is 85%. Most categories have high accuracy and recall rate. Category “3” may have lower indicators due to insufficient learning samples, but the overall prediction performance of the model is better.

The KNN model predicts 49 groups correctly and 5 groups incorrectly. The classification report is shown in Table 8. The accuracy of KNN model is 91%, which is higher than RF model. The accuracy and recall rate of most categories are high, and the prediction performance of the model is good.

The comparison between the prediction results of the above two models and the OLGA results are shown in Fig. 13. The red dots, dark blue square dots, and blue-green triangle points are the results of OLGA, RF model and KNN model, respectively. It can be seen intuitively from the figure that the prediction results of the two machine learning algorithms are very close to the simulation results of OLGA.

4. Conclusions

This paper aims to establish a natural gas pipeline liquid loading prediction model based on machine learning. Compared with OLGA simulation, the established data-driven model not only improves calculation efficiency and reduces workload, but also can provide technical support for gas pipeline flow safety.

(1) The topography is characterized by the combination of the inclination angle of the upward pipe and the mileage of the upward pipe. Various combinations of working conditions with different flow rates, pipe diameters, water contents, outlet pressures, pipe inclination angles, and the corresponding mileages of the upward pipe are designed by OLGA simulator to obtain a total data set of 1710 samples.

(2) The supervised learning approach is selected and the performances of six different algorithms, including support vector machine (SVM), decision tree (DT), random forest (RF), artificial neural network (ANN), naive Bayesian classification (NBC), and K nearest neighbor algorithm (KNN) are evaluated by the *k*-fold cross-validation method. Eventually, the RF and KNN algorithms are selected by comparing the accuracy and AUC values.

(3) After parameter tuning of RF and KNN, the accuracy, recall rate, F1 value, and accuracy of the RF and KNN are all above 90%, and the AUC of the RF and KNN are 0.99 and 0.96, respectively. In addition, RF and KNN are used to predict the location of the first liquid loading in gas pipeline, and the KNN algorithm performs better with an accuracy of 91%.

(4) Due to the lack of field data, the data in this paper is taken from OLGA simulations, so the sample amount is limited. Although we already have 1710 sets of data, after trying, these data are not enough to achieve accurate prediction of the amount of liquid loading. Machine learning methods depend on the quality and number of data sets. Therefore, a large amount of field data will be introduced in subsequent research to further improve the accuracy and application scope of the prediction model.

(5) The deviation of mechanism model calculation and the risk of over fitting in data-driven make the dual-driven model a more ideal solution. In future research, we will explore the hybrid modeling based on mechanism data driven to predict liquid loading.

Declaration of competing interest

There are no conflicts of interest to declare.

Acknowledgement

This work was supported by the National Science and

Technology Major Project of China (2016ZX05066005-001), Zhejiang Province Key Research and Development Plan (2021C03152) and Zhoushan Science and Technology Project (2021C21011), all of which are gratefully acknowledged.

References

Abubakar, A., Al-Wahaibi, Y., Al-Wahaibi, T., Al-Hashmi, A.R., Al-Ajmi, A., Eshrati, M., 2018. Effect of pipe diameter on horizontal oil-water flow before and after addition of drag-reducing polymer part II: holdup and slip ratio. J. Petrol. Sci. Eng. 162, 143–149. https://doi.org/10.1016/j.petrol.2017.12.015.

Andrejiova, M., Grincova, A., 2018. Classification of impact damage on a rubber-textile conveyor belt using Naïve-Bayes methodology. Wear 414–415, 59–67. https://doi.org/10.1016/j.wear.2018.08.001.

Barnea, D., 1987. A unified model for predicting flow-pattern transitions for the whole range of pipe inclinations. Int. J. Multiphas. Flow 13, 1–12. https://doi.org/10.1016/0301-9322(87)90002-4.

Brown, S.D., Myles, A.J., 2020. Decision tree modeling. Compr. Chemom. 625–659. https://doi.org/10.1016/B978-0-12-409547-2.00653-3.

Chen, S., Gong, J., Li, W., Yang, Q., Shi, G., Li, X., Shi, B., Song, S., Lv, P., Fan, D., Duan, X., 2021a. A new transient model of multi-scale bubble migration and evolution during gas-liquid flow in pipelines. J. Petrol. Sci. Eng. 205, 108888. https://doi.org/10.1016/j.petrol.2021.108888.

Chen, S., Gong, J., Li, W., Yang, Q., Shi, G., Li, X., Shi, B., Song, S., Lv, P., Fan, D., Duan, X., 2021b. A new transient model of multi-scale bubble migration and evolution during gas-liquid flow in pipelines. J. Petrol. Sci. Eng. 205, 108888. https://doi.org/10.1016/j.petrol.2021.108888.

Dong, Y., Ma, X., Fu, T., 2021. Electrical load forecasting: a deep learning approach based on K-nearest neighbors. Appl. Soft Comput. 99, 106900. https://doi.org/10.1016/j.asoc.2020.106900.

Fan, D., Gong, J., Zhang, S., Shi, G., Kang, Q., Xiao, Y., Wu, C., 2021. A transient composition tracking method for natural gas pipe networks. Energy. https://doi.org/10.1016/j.energy.2020.119131.

Finch, B.K., Beck, A.N., 2011. Socio-economic status and z-score standardized height-for-age of U.S.-born children (ages 2–6). Econ. Hum. Biol. 9, 272–276. https://doi.org/10.1016/j.ehb.2011.02.005.

Hashemizadeh, A., Maaref, A., Shateri, M., Larestani, A., Hemmati-Sarapardeh, A., 2021. Experimental measurement and modeling of water-based drilling mud density using adaptive boosting decision tree, support vector machine, and K-nearest neighbors: a case study from the South Pars gas field. J. Petrol. Sci. Eng. 109132. https://doi.org/10.1016/j.petrole.2021.109132.

He, G., Chen, D., Liao, K., Sun, J., Nie, S., 2019. A methodology for the optimal design of gathering pipeline system in old oilfield during its phased development process. Comput. Ind. Eng. 130, 14–34. https://doi.org/10.1016/j.cie.2019.02.016.

He, G., Tang, D., Yin, B., Sun, L., Ding, D., Liang, Y., Liao, K., 2018. Comparison and analysis of drainage measures for draining accumulated water condensed from wet CBM and transported in surface gathering pipeline network. J. Nat. Gas Sci. Eng. 56, 281–298. https://doi.org/10.1016/j.jngse.2018.06.017.

Hong, B., Li, X., Di, G., Li, Y., Liu, X., Chen, S., Gong, J., 2019. An integrated MILP method for gathering pipeline networks considering hydraulic characteristics. Chem. Eng. Res. Des. https://doi.org/10.1016/j.cherd.2019.08.013.

Hong, B., Li, X., Di, G., Song, S., Yu, W., Chen, S., Li, Y., Gong, J., 2020a. An integrated MILP model for optimal planning of multi-period onshore gas field gathering pipeline system. Comput. Ind. Eng. https://doi.org/10.1016/j.cie.2020.106479.

Hong, B., Li, X., Song, S., Chen, S., Zhao, C., Gong, J., 2020b. Optimal planning and modular infrastructure dynamic allocation for shale gas production. Appl. Energy 261, 114439. https://doi.org/10.1016/j.apenergy.2019.114439.

Huo, W., Li, W., Zhang, Z., Sun, C., Zhou, F., Gong, G., 2021. Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection. Energy Convers. Manag. 243, 114367. https://doi.org/10.1016/j.enconman.2021.114367.

Huo, Y., Bouffard, F., Joós, G., 2021. Decision tree-based optimization for flexibility management for sustainable energy microgrids. Appl. Energy 290, 116772. https://doi.org/10.1016/j.apenergy.2021.116772.

Izwan Ismail, A.S., Ismail, I., Zoveidavianpoor, M., Mohsin, R., Piroozian, A., Misnan, M.S., Sariman, M.Z., 2015. Experimental investigation of oil–water two-phase flow in horizontal pipes: pressure losses, liquid holdup and flow patterns. J. Petrol. Sci. Eng. 127, 409–420. https://doi.org/10.1016/j.petrol.2015.01.038.

Jain, B., Ranawat, N., Chittora, P., Chakrabarti, P., Poddar, S., 2021. A machine learning perspective: to analyze diabetes. Mater. Today Proc. https://doi.org/10.1016/j.matpr.2020.12.445.

Kanin, E.A., Osiptsov, A.A., Vainshtein, A.L., Burnaev, E.V., 2019. A predictive model for steady-state multiphase pipe flow: machine learning on lab data. J. Petrol. Sci. Eng. 180, 727–746. https://doi.org/10.1016/j.petrol.2019.05.055.

Kesana, N.R., Ibarra, R., Langsholt, M., Skartlien, R., Skjæraasen, O., Tutkun, M., 2018. Measurements of local droplet velocities in horizontal gas-liquid pipe flow with low liquid loading. J. Petrol. Sci. Eng. 170, 184–196. https://doi.org/10.1016/j.petrol.2018.06.019.

Khajenezhad, A., Bashiri, M.A., Beigy, H., 2021. A distributed density estimation algorithm and its application to naive Bayes classification. Appl. Soft Comput. 98, 106837. https://doi.org/10.1016/j.asoc.2020.106837.

Khaledi, H.A., Smith, I.E., Unander, T.E., Nossen, J., 2014. Investigation of two-phase flow pattern, liquid holdup and pressure drop in viscous oil–gas flow. Int. J. Multiphas. Flow 67, 37–51. https://doi.org/10.1016/j.ijmultiphaseflow.2014.07.006.

Lee, S., 2021. Monte Carlo simulation using support vector machine and kernel density for failure probability estimation. Reliab. Eng. Syst. Saf. 209, 107481. https://doi.org/10.1016/j.ress.2021.107481.

Liang, F., Hang, Y., Yu, H., Gao, Jifeng, 2021. Identification of gas-liquid two-phase flow patterns in a horizontal pipe based on ultrasonic echoes and RBF neural network. Flow Meas. Instrum. 79, 101960. https://doi.org/10.1016/j.flowmeasinst.2021.101960.

Lin, Z., Liu, X., Lao, L., Liu, H., 2020. Prediction of two-phase flow patterns in upward inclined pipes via deep learning. Energy 210, 118541. https://doi.org/10.1016/j.energy.2020.118541.

Mask, G., Wu, X., Ling, K., 2019. An improved model for gas-liquid flow pattern prediction based on machine learning. J. Petrol. Sci. Eng. 183, 106370. https://doi.org/10.1016/j.petrol.2019.106370.

Ming, R., He, H., Hu, Q., 2018. A new model for improving the prediction of liquid loading in horizontal gas wells. J. Nat. Gas Sci. Eng. 56, 258–265. https://doi.org/10.1016/j.jngse.2018.06.003.

Ong, C.L., Thome, J.R., 2011. Macro-to-microchannel transition in two-phase flow: Part 1 – two-phase flow patterns and film thickness measurements. Exp. Therm. Fluid Sci. 35, 37–47. https://doi.org/10.1016/j.expthermflusci.2010.08.004.

Qi, C., Fourie, A., Chen, Q., 2018. Neural network and particle swarm optimization for predicting the unconfined compressive strength of cemented paste backfill. Construct. Build. Mater. 159, 473–478. https://doi.org/10.1016/j.conbuildmat.2017.11.006.

Rodrigues, H.T., Pereyra, E., Sarica, C., 2018. A model for the thin film friction factor in near-horizontal stratified-annular transition two-phase low liquid loading flow. Int. J. Multiphas. Flow 102, 29–37. https://doi.org/10.1016/j.ijmultiphaseflow.2018.01.017.

Rodrigues, H.T., Soedarmo, A., Pereyra, E., Sarica, C., 2020. Droplet entrainment measurements under high-pressure two-phase low-liquid loading flow in slightly inclined pipes. J. Petrol. Sci. Eng. 187, 106767. https://doi.org/10.1016/j.petrol.2019.106767.

Salubi, V., Mahon, R., Oluyemi, G., Oyeneyin, B., 2021. Effect of two-phase gas-liquid flow patterns on cuttings transport efficiency. J. Petrol. Sci. Eng. 109281. https://doi.org/10.1016/j.petrol.2021.109281.

Saud, S., Jamil, B., Upadhyay, Y., Irshad, K., 2020. Performance improvement of empirical models for estimation of global solar radiation in India: a *k*-fold cross-validation approach. Sustain. Energy Technol. Assessments 40, 100768. https://doi.org/10.1016/j.seta.2020.100768.

Shi, B., Song, S., Chen, Y., Duan, X., Liao, Q., Fu, S., Liu, L., Sui, J., Jia, J., Liu, H., Zhu, Y., Song, C., Lin, D., Wang, T., Wang, J., Yao, H., Gong, J., 2021. Status of natural gas hydrate flow assurance research in China: a review. Energy Fuels 35, 3611–3658. https://doi.org/10.1021/acs.energyfuels.0c04209.

Shi, G., Fan, D., Gong, J., 2020. A new transient simulation method of natural gas-condensate two-phase flow in pipeline network. Chem. Eng. Sci. 223, 115742. https://doi.org/10.1016/j.ces.2020.115742.

Shi, G., Song, S., Shi, B., Gong, J., Chen, D., 2021. A new transient model for hydrate slurry flow in oil-dominated flowlines. J. Petrol. Sci. Eng. 196, 108003. https://doi.org/10.1016/j.petrol.2020.108003.

Shi, L., Zhang, S., Arshad, A., Hu, Y., He, Y., Yan, Y., 2021. Thermo-physical properties prediction of carbon-based magnetic nanofluids based on an artificial neural network. Renew. Sustain. Energy Rev. 149, 111341. https://doi.org/10.1016/j.rser.2021.111341.

Si, J., Wang, G., Li, P., Mi, J., 2021. A new skeletal mechanism for simulating MILD combustion optimized using Artificial Neural Network. Energy 237, 121603. https://doi.org/10.1016/j.energy.2021.121603.

Taitel, Y., Dukler, A.E., 1976. A model for predicting flow regime transitions in horizontal and near horizontal gas-liquid flow. AIChE J. 22, 47–55. https://doi.org/10.1002/aic.690220105.

Tiwary, A.K., Ghosh, S., Singh, R., Mukherjee, D.P., Shankar, B.U., Dash, P.S., 2020. Automated coal petrography using random forest. Int. J. Coal Geol. 232, 103629. https://doi.org/10.1016/j.coal.2020.103629.

Vieira, C., Stanko, M., Oplt, T., 2021. An improved model for predicting liquid loading onset in inclined pipes with non-uniform liquid wall films. J. Nat. Gas Sci. Eng. https://doi.org/10.1016/j.jngse.2021.103902.

Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., Rajan, A.R., 2021. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. Comput. Electr. Eng. 92, 107186. https://doi.org/10.1016/j.compeleceng.2021.107186.

Zhang, H., Shi, Y., Yang, X., Zhou, R., 2021. A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance. Res. Int. Bus. Finance 58, 101482. https://doi.org/10.1016/j.ribaf.2021.101482.

Zhang, H.Q., Wang, Q., Sarica, C., Brill, J.P., 2004. Unified model for gas-liquid pipe flow via slug dynamics - Part 1: model development. SPE Repr. Ser. 44–51. https://doi.org/10.1115/1.1615246.