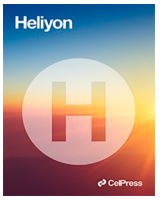
Heliyon 9 (2023) e17639

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



A decision tree model for accurate prediction of sand erosion in elbow geometry

Fahd Saeed Alakbari [a,b,**], Mysara Eissa Mohyaldinn [a,b,*], Mohammed Abdalla Ayoub [a,b], Abdullah Abduljabbar Salih [a,b], Azza Hashim Abbas [c]

[a] Petroleum Engineering Department, Universiti Teknologi PETRONAS, 32610, Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia
[b] Institute of Hydrocarbon Recovery, Universiti Teknologi PETRONAS, 32610, Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia
[c] School of Mining and Geosciences, Nazarbayev University, Nur Sultan, 010000, Kazakhstan

ARTICLE INFO

Keywords:
Erosion rate
Sand production
Decision tree
Data-driven methods
Machine learning

ABSTRACT

Erosion of piping components, e.g., elbows, is a hazardous phenomenon that frequently occurs due to sand flow with fluids during petroleum production. Early prediction of the sand's erosion rate (ER) is essential for ensuring a safe flow process and material integrity. Some models have been applied to determine the ER of the sand in the literature. However, these models have been created based on specific data to require a model for application to wide-range data. Moreover, the previous models have not studied relationships between independent and dependent variables. Thus, this research aims to use machine learning techniques, namely linear regression and decision tree (DT), to predict the ER robustly. The optimum model, the DT model, was evaluated using various trend analysis and statistical error analyses (SEA) techniques, namely the correlation coefficient (R). The evaluation results proved proper physical behavior for all independent variables, along with high accuracy and the DT model robustness. The proposed DT method can accurately predict the ER with R of 0.9975, 0.9911, 0.9761, and 0.9908, AAPRE of 5.0%, 6.27%, 6.26%, and 5.5%, RMSE of 2.492E-05, 6.189E-05, 9.310E-05, and 5.339E-05, and STD of 13.44, 6.66, 8.01, and 11.44 for the training, validation, testing, and whole datasets, respectively. Hence, this study delivers an effective, robust, accurate, and fast prediction tool for ER determination, significantly saving the petroleum industry's cost and time.

1. Introduction

Sand erosion is a critical dilemma in the petroleum industry. Erosion causes equipment damage, thereby producing environmental risks, decreasing the equipment lifespan, and increasing the maintenance cost of the equipment [1]. Therefore, it is necessary to predict erosion early before this issue becomes complicated and expensive.

Some scholars have attempted to determine erosion using various approaches. Jordan [2] introduced a model to assess erosion using multiphase flow based on Shirazi's [3] model. One limitation of these models is that they do not consider liquid holdup. A

* Corresponding author. Petroleum Engineering Department, Universiti Teknologi PETRONAS, 32610, Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia.
** Corresponding author. Petroleum Engineering Department, Universiti Teknologi PETRONAS, 32610, Bandar Seri Iskandar, Perak Darul Ridzuan, Malaysia.
E-mail addresses: fahd_19001032@utp.edu, alakbarifahd@gmail.com (F.S. Alakbari), mysara.eissa@utp.edu.my (M.E. Mohyaldinn).

https://doi.org/10.1016/j.heliyon.2023.e17639
Received 14 November 2022; Received in revised form 23 May 2023; Accepted 23 June 2023
Available online 25 June 2023

Nomenclature

ER	erosion rate, mm/kg
V_{SL}	superficial liquid velocity, m/s
V_{SG}	superficial gas velocity, m/s
D	pipe diameter, m.
d_p	particle size, μm
μ_L	liquid viscosity, (cP)
CFD	computational fluid dynamics
DIM	direct impingement model
DT	decision tree
PVT	pressure-volume-temperature
PB	physical behavior
cP	centipoise.
SEA	statistical error analyses
TA	trend analysis
APRE	average-percent-relative-error
R	correlation coefficient
AAPRE	average-absolute-percent-relative-error
RMSE	root-mean square-error
E_{max}	maximum-absolute-percent-relative-error
SD	Standard-deviation
E_{min}	minimum absolute-percent-relative-error

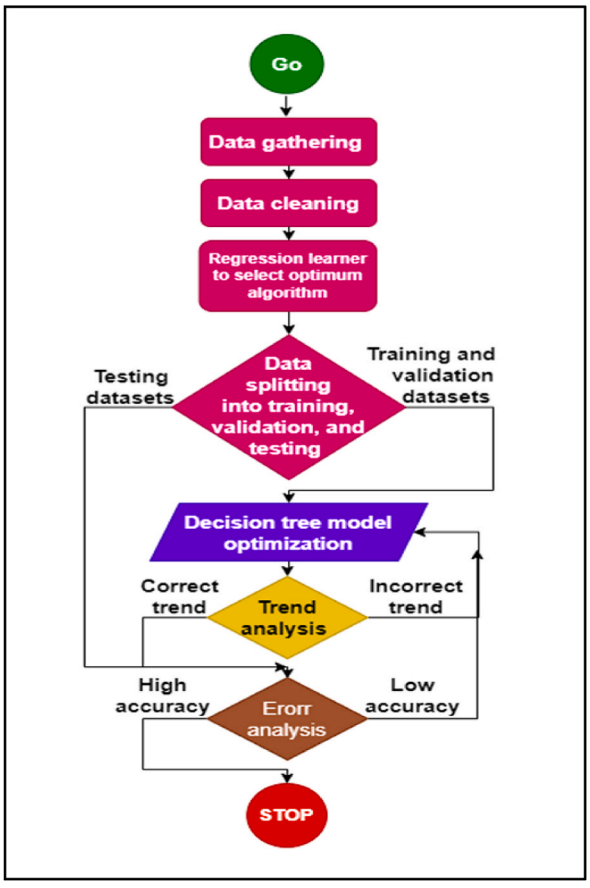


Fig. 1. Flowchart of the proposed scheme.

Table 1
The gathered data's statistical description.

Parameter	Pipe diameter (D), (m)	Particle size (d_p), (μm)	Liquid viscosity (μ_L), (cP)	Superficial liquid velocity (V_{SL}), (m/s)	Superficial gas velocity (V_{SG}), (m/s)	Erosion rate (ER), (mm/kg)
Minimum	0.0254	20.00	0.00	0.00	3.50	3.7×10^{-6}
Maximum	0.1016	550.00	10.00	6.20	222.00	0.26200
Mean	0.0715	260.66	1.07	0.29	34.57	0.01289
Standard Error	0.0012	6.00	0.10	0.04	1.67	0.00214
Median	0.0762	300.00	1.00	0.04	27.00	0.00038
Mode	0.0762	300.00	1.00	0.00	15.00	0.00146
Standard Deviation	0.0242	117.63	2.04	0.82	32.77	0.04203
Sample Variance	0.0006	13836.44	4.14	0.68	1073.83	0.00177
Kurtosis	−0.6806	0.67	14.69	28.95	10.28	17.4033
Skewness	−0.4687	0.49	3.94	5.23	3.05	4.15241
Range	0.0762	530.00	10.00	6.20	218.50	0.26200

computational-fluid-dynamics (CFD) method was utilized to obtain the sand's erosion rate (ER). However, the CFD method has shortcomings: it uses only gas production systems [4]. Mohyaldinn et al. [5] created a model based on Salama's [6] model to find ER. However, their model was used in pure gas and high and low gas-liquid ratio conditions. A random-forest regression model was utilized to obtain the erosion in the elbows. However, this method is limited in predicting specific liquid-dominant flow regimes [7]. A probability model was used to forecast ER in elbows under annular flow. However, the probability model cannot be used for other multiphase-flow patterns [8].

In the 2020s, the direct impingement model (DIM) was utilized to determine the ER, indicating that the model has the Salama [6] model's simplicity and the DIM model's accuracy [9]. Zhang and Xu [10] developed a least-squares-boosting model to determine the ER and applied it to single and multiphase-flow patterns. A Gaussian process regression was utilized to obtain the ER; however, the model has limitations. The model was based on specific datasets and had a relative error of approximately 20% [11]. Therefore, ER prediction models are required to obtain better accuracy and to improve the model performance using a wide range of datasets. Furthermore, previous models have not proved the proper physical behavior (PB).

A decision tree (DT) is a non-parametric supervised learning technique applied for classification or regression in this study [12]. More details about the DT approach are discussed in the following subsection, i.e., 2.2 Decision tree method. The DT model has been successfully used in petroleum engineering applications. Almashan et al. [13] used the DT method to determine pressure-volume-temperature properties. The liquid holdup in the two-phase gas and liquid flow was determined using the DT method [14]. The multiphase flow parameters, namely pressure drop, were determined using the DT method [15]. Drilling fluid loss was predicted using the DT approach [16]. The DT method is used to find drilling lithology rocks by applying sonic data [17]. However, previous models utilized the DT method without studying the PB. Therefore, this study uses DT with trend analysis (TA) to show the PB.

The novelty of this research lies in developing a DT model to determine the ER robustly and accurately. Therefore, datasets were gathered from various references to utilize a wide range to represent a robust model. The developed DT model was evaluated using various techniques to demonstrate a robust and accurate proposed method. The datasets were divided into three subsections, training, validation, and testing, for overcoming overfitting and underfitting issues. Several statistical error analyses (SEA), that is, the absolute-average-percentage-relative-error (AAPRE), correlation coefficient (R), average-percentage-relative-error (APRE), standard-deviation (SD), maximum-absolute-percent-relative-error (E_{max}), root-mean-square-error (RMSE), and minimum-absolute-percent-relative-error (E_{min}) have been used to assess and demonstrate the DT model's accuracy. Moreover, TA can be utilized to assess the proposed model and to prove proper PB.

2. Methodology

This study was conducted following the stages demonstrated in Fig. 1. Initially, the datasets were gathered from different studies to obtain a wide range of datasets. Subsequently, the collected datasets were cleaned for use as actual datasets and to increase the models' performance. Then, the cleaned datasets were used to apply regression learner to select the optimum algorithm to determine the sand's erosion rate (ER). Then, the cleaned data were split to three subdivisions: testing, validation, and training to train the optimum algorithm: decision tree (DT) to determine the ER. The validation and training datasets were used to train and validate the model. The training and validation datasets were shuffled to ensure the model did not have any overfitting or underfitting. The testing dataset was used to compare all the models with the same dataset to make a fair comparison between the models. In addition, all datasets were computed to demonstrate the DT model's performance. After dividing the datasets, the proposed DT model was optimized to determine the ER robustly and accurately. Trend analyses were performed to show the correct PB. After all independent variables followed the precise trends, the proposed DT model's accuracy was checked using some SEA, namely RMSE and R.

Table 2
Samples of the gathered data.

No.	D, (m)	d_p, (μm)	μ_L, (cP)	V_{SL}, (m/s)	V_{SG}, (m/s)	ER, (mm/kg)
1	0.0762	300	1	0.37	27.2	0.00037
2	0.0762	300	1	0.55	27.4	0.000191
3	0.0762	300	1	0.47	31.1	0.000242
4	0.0762	300	1	0.46	49	0.000743
5	0.0762	20	10	0.42	27.2	0.00000965
6	0.0762	150	10	0.53	10.8	0.00000369
7	0.0762	150	10	0.5	18.5	0.0000151
8	0.0762	150	10	0.3	27.2	0.000146
9	0.0762	300	10	0.44	11.2	0.0000145
10	0.0762	300	10	0.36	18.2	0.0000708
11	0.0762	300	10	0.29	27.3	0.000291
12	0.0762	300	10	0.32	27.3	0.000281

2.1. Data collection and pre-processing

Collecting numerous data to build a model is time-consuming and difficult. However, in this study, 384 datasets were collected from various sources [4,7,18–26] to use wide data ranges. The ranges of the collected datasets are listed in Table 1. Table 2 shows samples of the collected datasets. Fig. 2(a–f) displays the histograms of the parameters for the gathered datasets.

Poor quality data cause substantial problems in developing machine-learning models [27]. Therefore, data cleaning plays a crucial role, pre-empting the presence of duplicated and corrupt datasets to build the proposed models. Some of the collected datasets had different output values for the same input values because the experimental datasets were measured more than once without considering the average. Therefore, the average of the outputs for the same input values was determined to obtain one output for each input parameter. Subsequently, the outliers of the collected datasets were detected and removed using a box and whisker plots were clarified in detail [28]. The box and whisker plot was established by Tukey [29]. The box and whisker consist of some parameters, namely interquartile-range (IQR), as shown in Fig. 3. Any values less than the lower boundary or higher than the upper boundary can be the outliers [30].

After the outliers of the collected datasets were removed, the particle size and liquid viscosity were 300 μm and 1 cP (cP) for all datasets. The particle size and liquid viscosity were constant in the clean datasets; therefore, these parameters were removed from the clean datasets. 72% of the total collected datasets were removed because of many duplicated and corrupt datasets in the gathered datasets. The clean data comprised 106 datasets without duplicated and/or corrupt datasets. Statistical descriptions of the clean datasets are presented in Table 3. Some samples of the clean data are shown in Table 4. However, the histograms of the parameters for the clean datasets are demonstrated in Fig. 4(a–f). The various studies [4,7,18–26] considered pipe diameter, particle size, liquid viscosity, and superficial liquid and gas velocities as inputs to determine the erosion rate. In this study, the same input parameters are considered to predict the erosion rate; however, the particle size and liquid viscosity were constant in the clean datasets; therefore, these parameters were not used to predict the erosion rate.

2.2. Machine learning techniques evaluation

Different machine learning approaches were applied to determine the ER. The methods are decision tree with small leaf size, ensemble boosted trees, interactions linear regression, stepwise linear regression, simple linear regression, robust linear regression, support vector machine (SVM) with Gaussian kernel function and 1.7 kernel scale, ensemble bagged trees, decision tree with medium leaf size, SVM with Gaussian kernel function and 0.43 kernel scale, SVM with linear-kernel-function, squared-exponential Gaussian-process-regression (GPR), rational-quadratic-GPR, SVM with Gaussian-kernel-function and 6.9 kernel-scale, Matern 5/2 GPR, decision tree with large leaf size, exponential GPR, SVM with three kernels polynomial order, and SVM with two kernels polynomial order. The different machine-learning methods were validated. After that, all machine learning methods were compared using RMSE, coefficient-of-determination (R^2), mean-square-error (MSE), and mean-absolute-error (MSE) to select which algorithm to use for determining the ER. The decision tree with a small leaf size and a minimum leaf size of 4 had the lowest error as the best model to obtain the ER. Then, the best model, the decision tree, was evaluated and studied in detail to accurately determine the ER with the proper relations between the independent variables and dependent variable to show the accurate PB and prove the robust model.

2.3. Decision tree technique

A decision tree (DT) can approximate discrete-valued target functions. The DT is expressed as a set of if-then rules [31]. It comprises parameters, including the root, decision, and leaf nodes, as shown in Fig. 5. In addition, the terms used in the DT method include splitting, subtree or branch, and pruning, which implies eliminating specific nodes, as shown in Fig. 5. Depending on the case used, a DT can be a regression or classification tree [32].

Decision trees categorize instances by arranging them in a bottom-up approach from the root to leaf nodes, indicating the categorization of the instances. Each node implies testing an instance attribute, and each subtree descends from the node to one of the

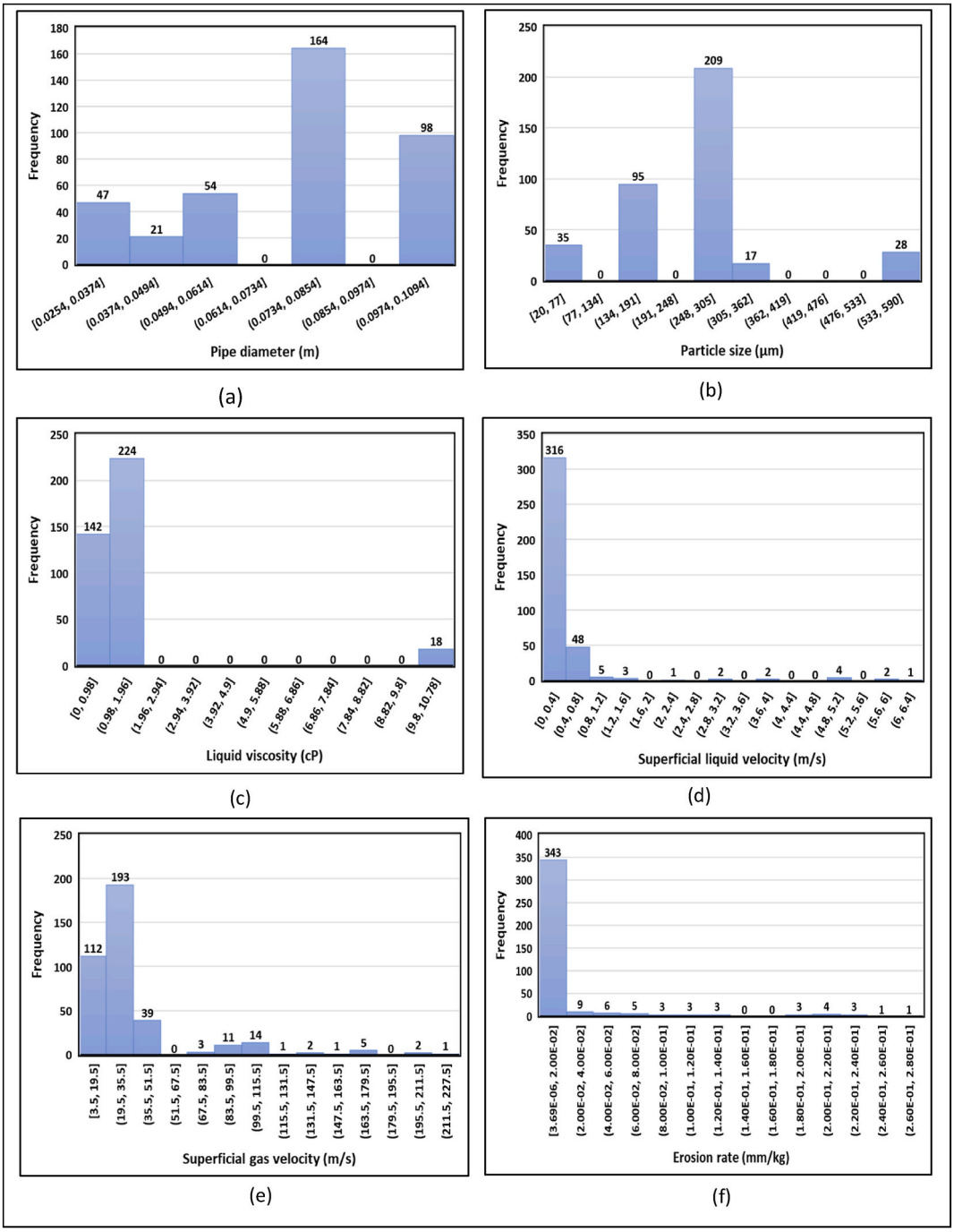


Fig. 2. Histograms of the parameters (a) pipe diameter, (b) particle size, (c) liquid viscosity, (d) superficial liquid velocity (e) superficial gas velocity, and (f) erosion rate for the gathered datasets.

possible values for this attribute. An instance can be categorized by beginning at the root node, checking the attribute identified by this node, and moving down the subtree corresponding to the attribute value in a specific example. This procedure is repeated for a subtree rooted at a new node [31].

The DT method is one of the most commonly applied techniques in prediction and has been used in different applications. Furthermore, the DT method is robust to noisy data [33,34]. It can automatically control missing values [35]. Another benefit of DT is that it effectively solves nonlinear problems [36]. The DT can use continuous and categorical variables as inputs to develop a model [37]. The DT technique can provide high performance [38].

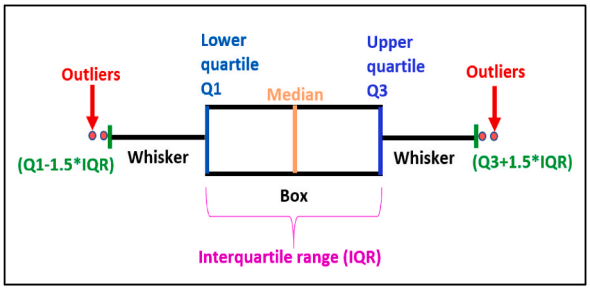


Fig. 3. The box and whisker plot.

Table 3
The clean data's statistical description.

Parameter	D, (m)	V_{SL}, (m/s)	V_{SG}, (m/s)	ER, (mm/kg)
Minimum	0.0762	0.0000	11.000	7.3×10^{-5}
Maximum	0.1016	0.1630	41.500	0.00178
Mean	0.0898	0.0744	27.052	0.00047
Standard Error	0.0011	0.0046	0.664	3.0×10^{-5}
Median	0.1016	0.0555	27.000	0.00042
Mode	0.1016	0.0400	36.000	0.00043
Standard Deviation	0.0124	0.0542	7.739	0.00035
Sample Variance	0.0002	0.0029	59.894	1.2×10^{-7}
Kurtosis	−1.9692	−1.2534	−0.806	2.39626
Skewness	−0.1490	0.4176	−0.100	1.60101
Range	0.0254	0.1630	30.500	0.00171

Table 4
Samples of the clean data.

No.	D, (m)	d_p, (μm)	μ_L, (cP)	V_{SL}, (m/s)	V_{SG}, (m/s)	ER, (mm/kg)
1	0.0762	300	1	0.02	15.2	0.000303
2	0.0762	300	1	0.01	15.2	0.000305
3	0.0762	300	1	0.01	15.2	0.000234
4	0.0762	300	1	0.1	11	7.98E-05
5	0.0762	300	1	0.09	27	0.00154
6	0.1016	300	1	0.15	15	0.000102
7	0.1016	300	1	0.1	22	0.000116
8	0.1016	300	1	0.01	23	0.000375
9	0.1016	300	1	0.01	21	0.000278
10	0.1016	300	1	0.02	22	0.000181
11	0.1016	300	1	0.02	22	0.00019
12	0.1016	300	1	0.04	23	0.000256

2.4. Trend analysis (TA)

The purpose of conducting a trend analysis (TA) is to assess how well a model can handle uncertainty. It involves examining the connections between independent and dependent variables in the model. By identifying errors in the models, TA reveals unexpected relationships between these variables, highlighting the importance of demonstrating the models' reliability. Additionally, TA helps identify and eliminate unnecessary components in the model structure [39]. Furthermore, TA has been used to identify significant links among observations, model inputs, and predictions, guiding the development of robust models [40]. Therefore, TA plays a crucial role in this study.

In this investigation, a specific independent parameter is selected for analysis, while other parameters are held constant at their mean values [41–43]. The proposed model determines the values of these parameters by modifying the studied independent variables. Subsequently, graphs are plotted with the independent parameter values on the x-axis and the dependent parameter values on the y-axis. The objective is to ensure that the relationship between the independent and dependent variables follows the correct behavior (PB). The independent variables chosen for TA in this study are pipe diameter (D), (m), superficial liquid velocity (VSL), (m/s), and superficial gas velocity (VSG), (m/s).

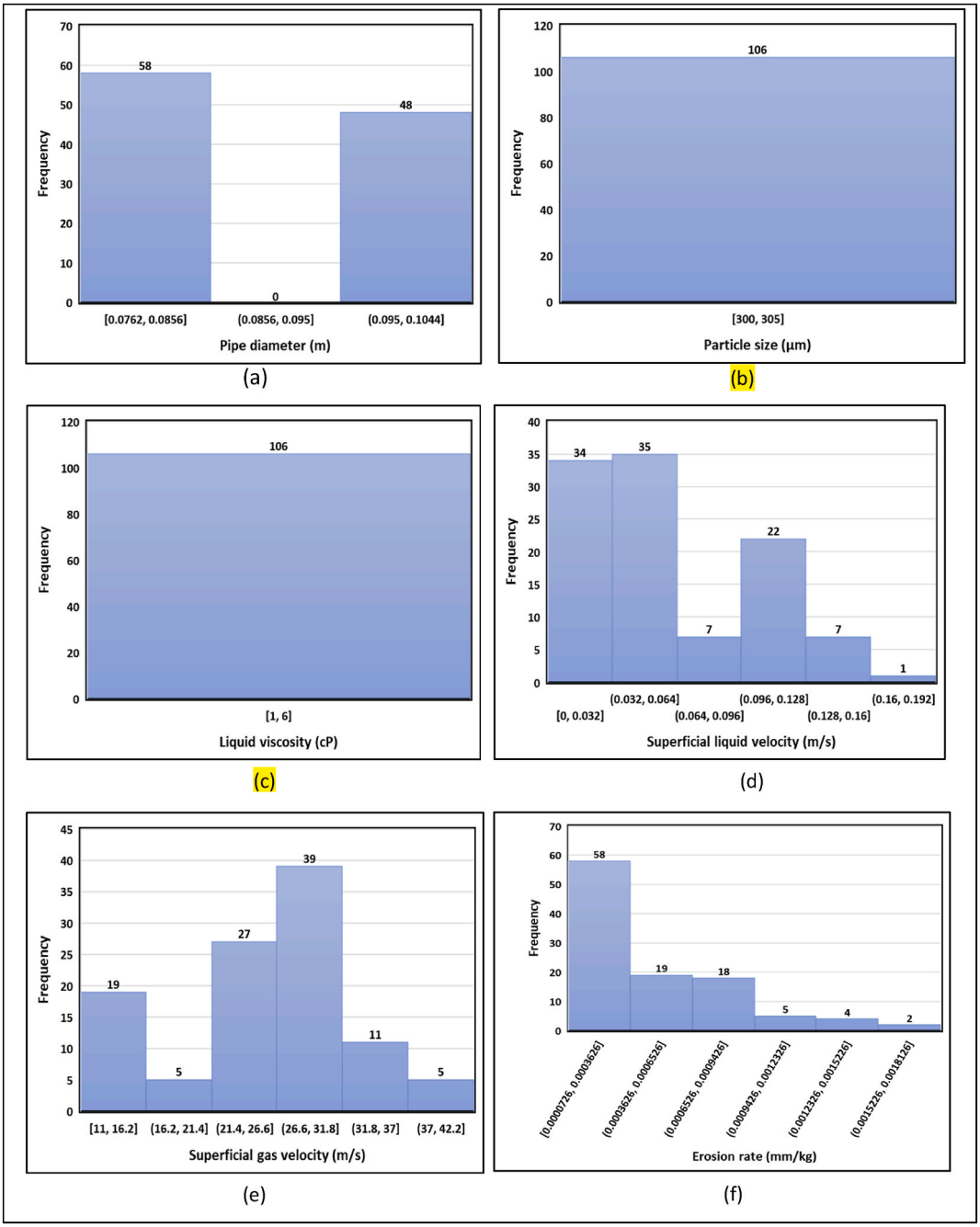


Fig. 4. Histograms of parameters (a) pipe diameter, (b) particle size, (c) liquid viscosity, (d) superficial liquid velocity (e) superficial gas velocity, and (f) erosion rate for clean datasets.

2.5. Statistical error analysis (SEA)

The SEA can be applied to present models' accuracy. In this study, different SEA are used: namely, APRE, RMSE, R, AAPRE, and SD. The SEA' equations are provided in the Supporting **information**.

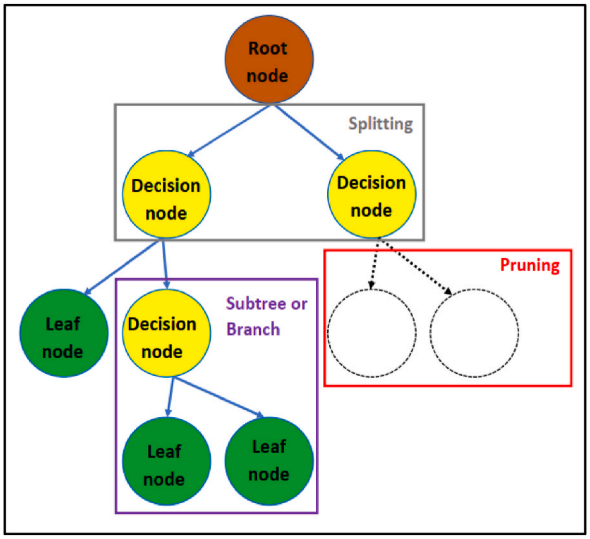


Fig. 5. Diagram of decision tree and its parameters and terms.

Table 5
Used parameters for machine learning methods.

No.	Model	Model Type
1	Decision tree with small leaf size	Minimum-leaf-size: 4; Maximum: 105; Merge leaves: on; Prune: on; Split criterion: mse; Surrogate-decision-splits: off.
2	Ensemble boosted trees	Learning-rate: 0.1; Minimum-leaf-size: 8; Method: tree boosting with least squares (LS); Number-of-learners: 30.
3	Interactions linear regression	Terms: interactions; Robust option: off.
4	Stepwise linear regression	Initial-terms: linear; Upper-bound-on-terms: interactions; Maximum-number-of-steps: 1000.
5	Simple linear regression	Robust option: off; Terms: linear.
6	Robust linear regression	Robust option: on; Terms: linear.
7	SVM with Gaussian kernel function and 1.7 kernel scale	Kernel-function: Gaussian; Standardize: true; Kernel-scale: 1.7; Box-constraint: 3.9585e-04; Epsilon: 3.9585e-05.
8	Ensemble bagged trees	Number-of-learners: 30; Minimum-leaf-size: 8; Learning-rate: 1.
9	Decision tree with medium leaf size	Surrogate-decision-splits: off; Minimum parent: 24; Maximum splits: 105; Merge leaves: on; Prune: on; Minimum-leaf-size: 12.
10	SVM with Gaussian kernel function and 0.43 kernel scale	Kernel-function: Gaussian; Standardize: true; Kernel-scale: 0.43; Box-constraint: 3.9585e-04; Epsilon: 3.9585e-05.
11	SVM with a linear kernel function	Kernel-function: linear; Standardize: true; Kernel-scale: automatic; Box constraint: 3.9585e-04; Epsilon: 3.9585e-05.
12	Squared exponential Gaussian process regression (GPR)	Basis-function: constant; Kernel-function: squared exponential; Signal-standard-deviation: automatic; Kernel-scale: automatic; Optimize-numeric-parameters: true; Standardize: true; Use-isotropic-kernel: true.
13	Rational quadratic GPR	Basis-function: constant; Signal-standard-deviation: automatic; Optimize-numeric-parameters: true; Sigma: automatic; Kernel-function: Rational-quadratic; Standardize: true; Use-isotropic-kernel: True; Kernel-scale: automatic.
14	SVM with Gaussian kernel function and 6.9 kernel scale	Kernel-function: Gaussian; Box-constraint: 3.9585e-04; Epsilon: 3.9585e-05; Kernel-scale: 6.9; Standardize: true.
15	Matern 5/2 GPR	Basis-function: Constant; Kernel-function: Matern 5/2; Signal-standard-deviation: automatic; Optimize-numeric-parameters: true; Standardize: true; Use isotropic-kernel: true; Sigma: automatic; Kernel-scale: automatic.
16	Decision tree with large leaf size	Minimum-leaf-size: 36; Surrogate-decision-splits: Off; Minimum parent: 72; Maximum splits: 105; Merge leaves: on; Prune: on.
17	Exponential GPR	Basis-function: constant; Sigma: automatic; Kernel-function: exponential; Optimize-numeric-parameters: true; Use isotropic-kernel: true; Kernel-scale: automatic; Standardize: true; Signal-standard deviation: automatic.
18	SVM with 3 kernel polynomial order	Kernel-function: polynomial; Standardize: true; Kernel-scale: automatic; Box constraint: 3.9585e-04; Epsilon: 3.9585e-05; Kernel polynomial order: 3.
19	SVM with 2 kernel polynomial order	Kernel-function: polynomial; Standardize: true; Kernel-scale: automatic; Box constraint: 3.9585e-04; Epsilon: 3.9585e-05; Kernel polynomial order: 2.

Table 6
Evaluation of the machine learning methods.

Rank	Model	RMSE	R^2	MSE	MAE
1	Decision tree with small leaf size	0.000150	0.850	2.25E-08	8.48E-05
2	Ensemble boosted trees	0.000152	0.850	2.31E-08	9.30E-05
3	Interactions linear regression	0.000205	0.730	4.19E-08	1.38E-04
4	Stepwise linear regression	0.000206	0.720	4.25E-08	1.38E-04
5	Simple linear regression	0.000223	0.670	4.98E-08	1.69E-04
6	Robust linear regression	0.000224	0.670	5.00E-08	1.69E-04
7	SVM with Gaussian kernel function and 1.7 kernel scale	0.000234	0.640	5.46E-08	1.72E-04
8	Ensemble bagged trees	0.000240	0.620	5.78E-08	1.69E-04
9	Decision tree with medium leaf size	0.000246	0.600	6.05E-08	1.67E-04
10	SVM with Gaussian kernel function and 0.43 kernel scale	0.000266	0.540	7.08E-08	2.17E-04
11	SVM with a linear kernel function	0.000275	0.500	7.59E-08	2.05E-04
12	Squared exponential Gaussian process regression (GPR)	0.000295	0.430	8.68E-08	2.22E-04
13	Rational quadratic GPR	0.000295	0.430	8.68E-08	2.22E-04
14	SVM with Gaussian kernel function and 6.9 kernel scale	0.000299	0.410	8.97E-08	2.43E-04
15	Matern 5/2 GPR	0.000312	0.360	9.74E-08	2.40E-04
16	Decision tree with large leaf size	0.000322	0.320	1.03E-07	2.26E-04
17	Exponential GPR	0.000391	0.000	1.53E-07	3.17E-04
18	SVM with 3 kernel polynomial order	0.000797	−3.160	6.35E-07	6.57E-04
19	SVM with 2 kernel polynomial order	0.000846	−3.690	7.16E-07	6.71E-04

Table 7
Used parameters for the DT model.

Parameter	Optimum value
Method	Tree
Type	regression
Minimum number of branch node observations	2
Min. leaf	1
Merge leaves	on
Prune	on
Split criterion	mse
Prune criterion	mse
Quadratic error tolerance	1e-6

3. Results and discussions

3.1. Machine learning methods' results

Table 5 shows the hyperparameters for the different machine learning methods to obtain the ER: decision tree with small leaf size, ensemble boosted trees, interactions linear regression, stepwise linear regression, simple linear regression, robust linear regression, SVM with Gaussian kernel function and 1.7 kernel scale, ensemble bagged trees, decision tree with medium leaf size, SVM with Gaussian kernel function and 0.43 kernel scale, SVM with linear-kernel-function, squared exponential-Gaussian-process regression (GPR), rational-quadratic-GPR, SVM with Gaussian-kernel-function and 6.9 kernel scale, matern 5/2 GPR, decision tree with large leaf size, exponential GPR, SVM with three kernels polynomial order, and SVM with two kernels polynomial order. The regression learner in MATLAB was applied to obtain optimum hyperparameters of all models to determine the ER.

After that, the different machine learning methods were evaluated using RMSE, R^2, MSE, and MAE [44]. Then, the machine learning methods to determine the CTD are ranked based on the low RMSE and high R^2, Table 6. The first rank model to determine the ER is the fine decision tree with RMSE of 0.000150 and R^2 of 0.850, while the second rank model is the ensemble boosted trees with RMSE of 0.000152 and R^2 of 0.850. The third rank model to obtain the ER is the interactions linear regression with RMSE of 0.000205 and R^2 of 0.730. The decision tree (DT) was chosen as the optimum ER model. Finally, the DR was evaluated to accurately determine the ER with the proper TA to demonstrate the proper relations between the independent and dependent variables and display the appropriate PB to prove the robust model.

3.2. Decision tree model

The proposed DT model was applied to determine the ER accurately and robustly using MATLAB software. The optimized parameters for the DT model are presented in Table 7. In this study, the fitted binary DT for regression was successfully applied to determine the ER with high accuracy. Tree inputs, D, V_{SL}, and V_{SG}, were used to predict the ER. Merge leaves "on" indicates that the models obtained from the similar parent node can merge and acquire a sum. The optimal sequence of pruned branches was subsequently obtained; however, it cannot prune the regression tree [45]. The algorithm that chooses the best predictor uses a separate predictor to maximize the split-criterion gain over all possible predictor splits [46]. Prune "on" implies that the model can grow the

Table 8
SEA of the DT model.

Datasets	APRE (%)	AAPRE (%)	E_{max} (%)	E_{min} (%)	RMSE	R	STD
Training	5.00	5.00	72.73	0	2.492E-05	0.9975	13.44
Validation	6.27	6.27	24.72	0	6.189E-05	0.9911	6.66
Testing	6.26	6.26	24.72	0	9.310E-05	0.9761	8.01
All	5.50	5.50	72.73	0	5.339E-05	0.9908	11.44

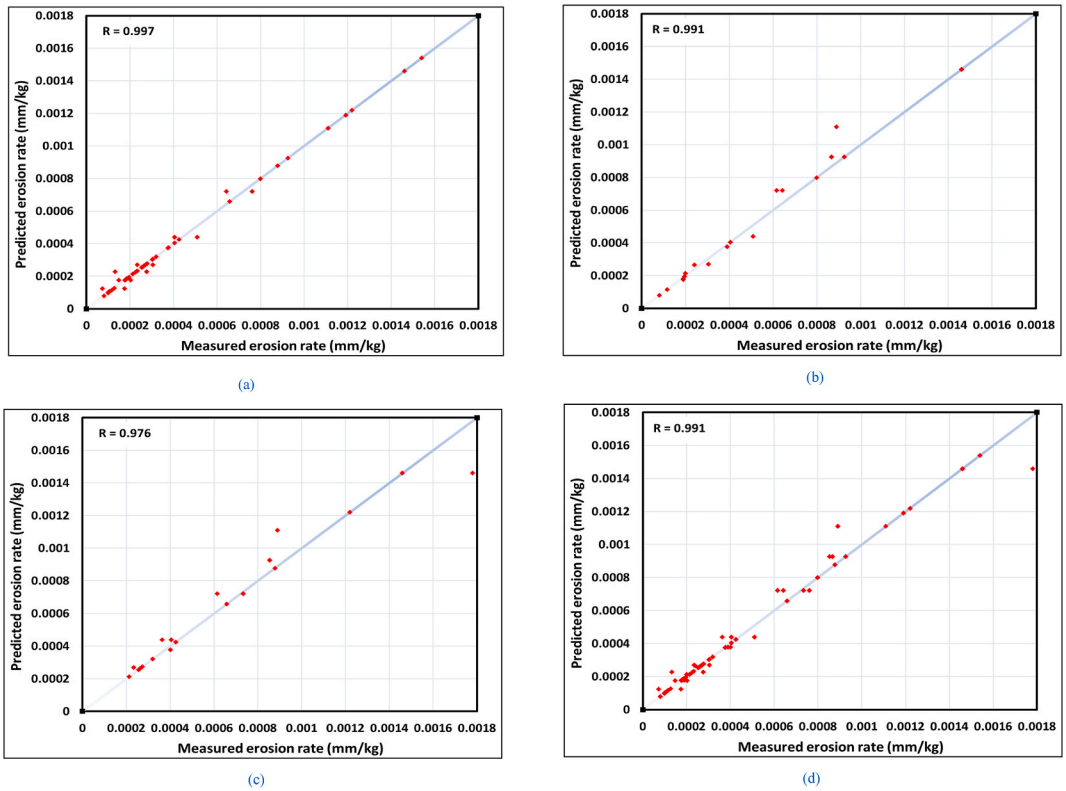


Fig. 6. (a). Cross-plot of the erosion rate for the training dataset. (b). Cross-plot of the erosion rate for the validation dataset. (c). Cross-plot of the erosion rate for the testing dataset. (d). Cross-plot of the erosion rate for the whole datasets.

regression tree and find the optimum sequence of pruned branches; however, it should not prune the regression tree [45–47].

Consequently, the proposed DT with these optimized parameters could accurately predict the ER. After the model was developed, it was evaluated using different methods, which will be discussed in the next section. The optimum hyperparameters of DT were chosen by applying the grid search method. The grid search is a process that systematically searches through a specified subset of the hyperparameter space of the targeted algorithm [48]. In the grid search technique, each parameter changes in their types or values while keeping the other fixed. After that, the DT model's accuracy and the proper trend analysis were checked. Finally, the best hyperparameters were chosen for the highest accuracy with the proper trend analysis, as shown in Table 7.

3.3. Decision tree model evaluation

The proposed DT model was evaluated using various procedures. The data was split into three subdivisions to solve overfitting and underfitting problems by checking SEA for the three subdivisions. Statistical-error analyses, namely RMSE, STD, R, and plots (i.e., error histograms), were achieved to represent high accuracy of the DT model for the training, validation, testing, and whole datasets. The closed SEA values of the three subdivisions indicate no overfitting or underfitting. However, the different SEA values show overfitting or underfitting issues. TA was performed to show proper PB.

3.3.1. Statistical error analysis

The different SEA, APRE, AAPRE, E_{min}, E_{max}, RMSE, STD, and R of the proposed DT model for all datasets, are shown in Table 8. The leading indicators in this study were AAPRE and R. The training, validation, testing, and whole datasets had AAPRE of 5%, 6.27%,

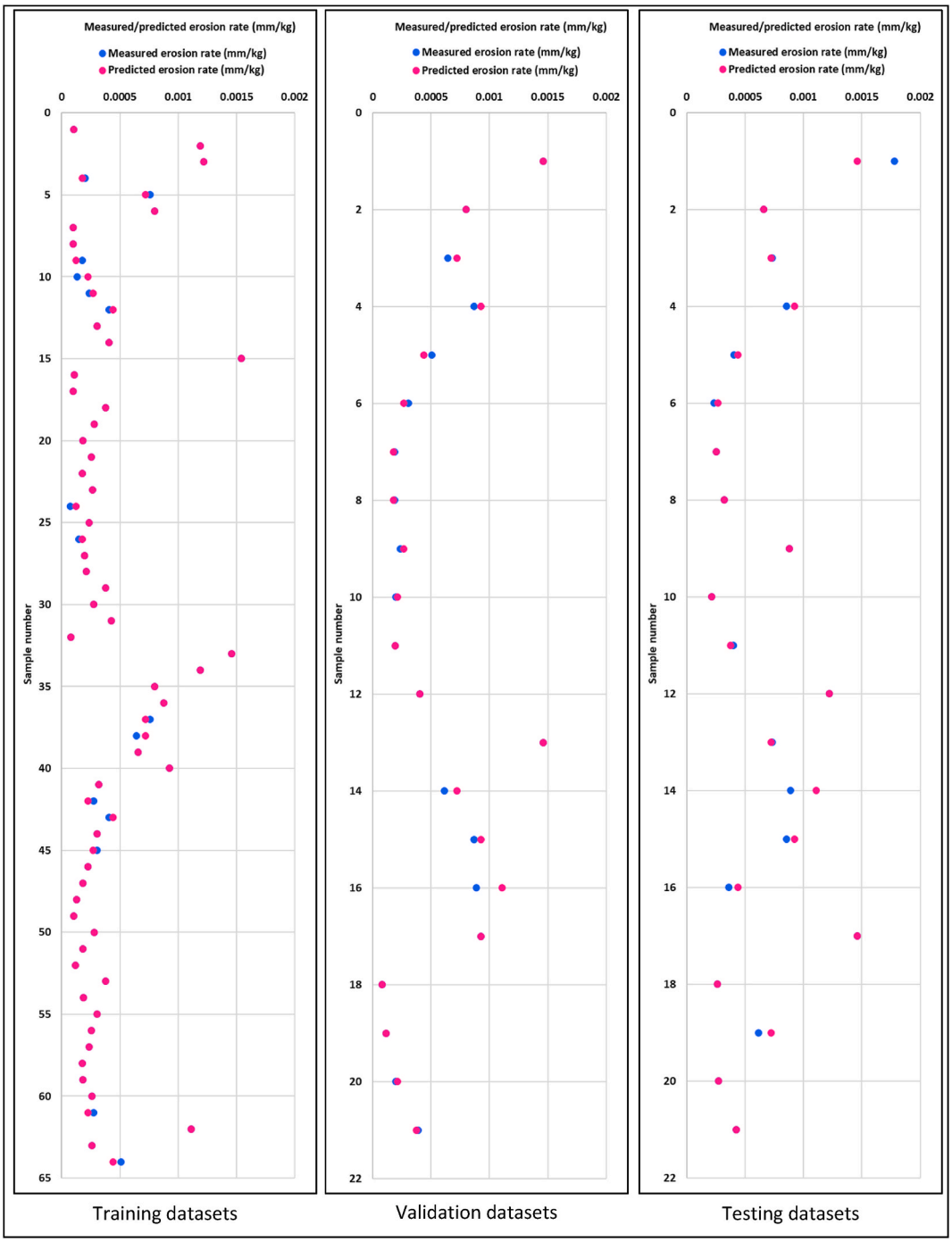


Fig. 7. Measured/predicted erosion rate (mm/kg) for training, validation, and testing datasets.

6.26%, and 5.50%, and R values of 0.9975, 0.9911, 0.9761, and 0.9908, respectively. Therefore, the training, validation, testing, and whole datasets have a low error, that is, AAPRE, proving that the DT model can accurately obtain the ER. In addition, all datasets have the closest AAPRE and R, confirming the robustness of the proposed DT model without any overfitting or underfitting problems.

3.3.2. Cross-plotting

Fig. 6(a–d) present the cross-plotting of the different datasets: training, validation, testing, and whole data. As shown in Fig. 6 (a), the cross-plotting of the training dataset implies that most data points are in a straight line, proving the accurate DT model for the training dataset. The cross-plotting for the validation, testing, and whole datasets are shown in Fig. 6(b–d). Most data points are in a

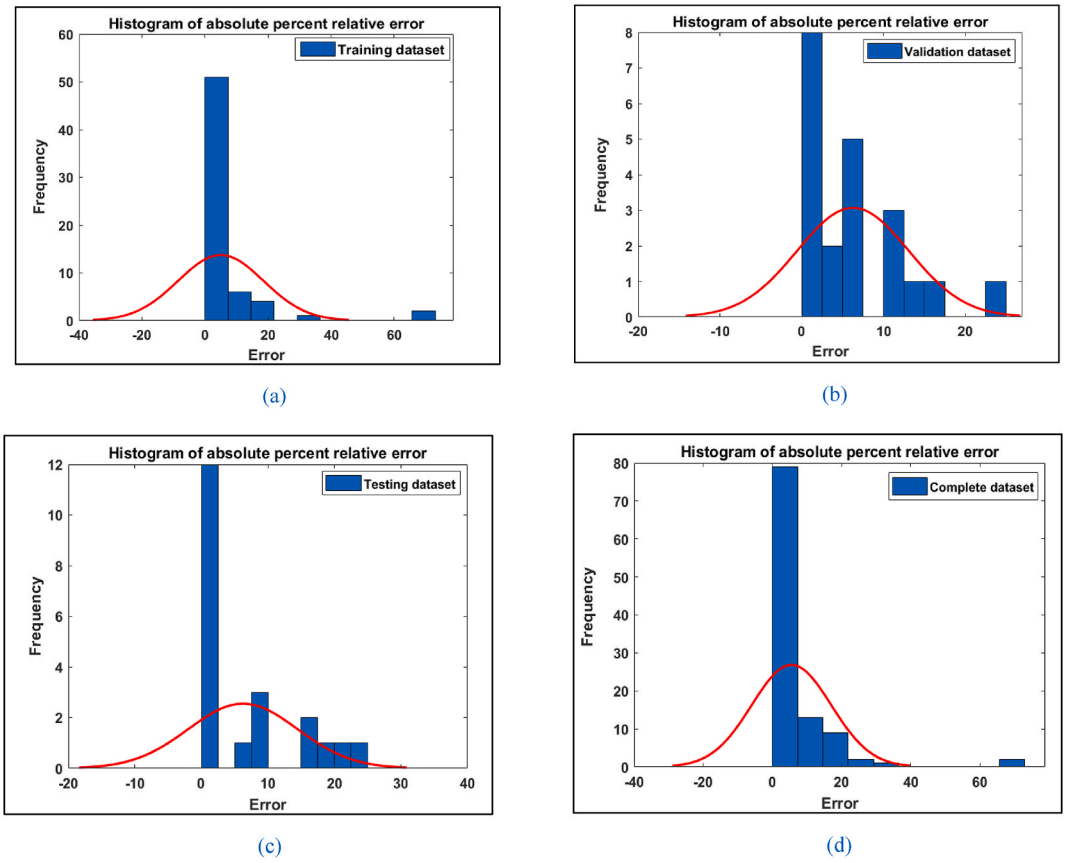


Fig. 8. (a). Histogram of erosion rate errors for the training dataset. (b). Histogram of erosion rate errors for the validation dataset. (c). Histogram of erosion rate errors for the testing dataset. (d). Histogram of erosion rate errors for the whole dataset.

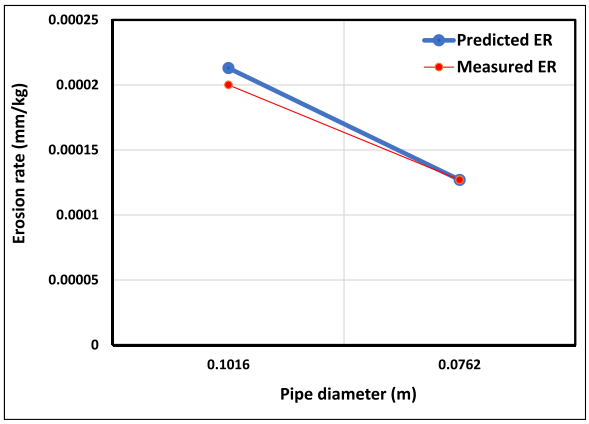


Fig. 9. Impact of pipe diameter on erosion rate.

straight line for the validation, testing, and entire datasets, Fig. 6(b–d), indicating the accuracy of the proposed DT model in determining the ER for these datasets. Furthermore, Fig. 6(a–d) show an excellent match between the actual and predicted values. In addition, the actual or measured and predicted ER data points for the different datasets, that are, training, validation, and testing datasets, are displayed in Fig. 7, demonstrating that the actual ER data points match the predicted ER data points for the different datasets, that are, training, validation, and testing datasets. The cross-plotting in Fig. 6(a–d) and measured/predicted ER plots in Fig. 7 prove that the DT model can robustly determine the ER without any over-fitting or under-fitting problems.

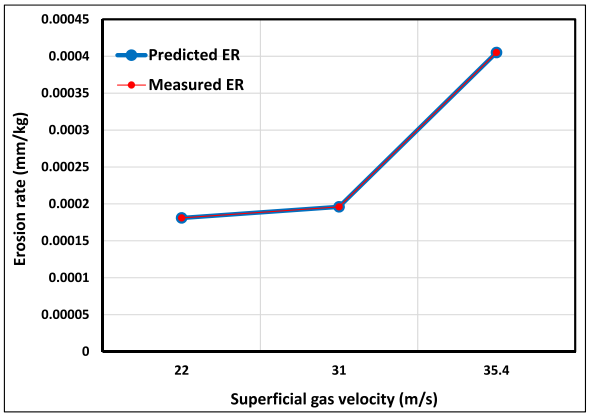


Fig. 10. Impact of superficial gas velocity on erosion rate.

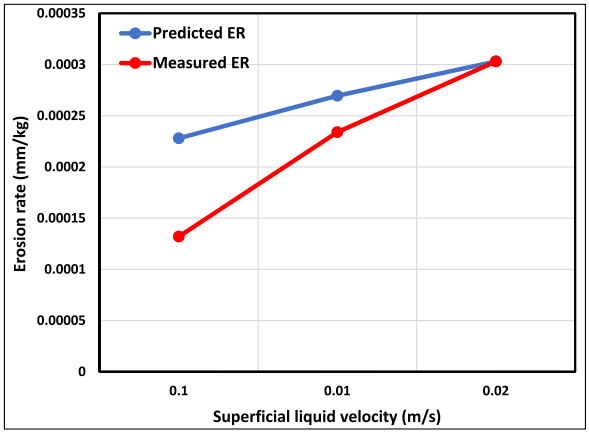


Fig. 11. Impact of superficial liquid velocity on erosion rate.

3.3.3. Error histograms

The absolute percent relative error histograms for the different datasets are shown in Fig. 8(a–d). As shown in Fig. 8(a–d), the training, validation, testing, and whole datasets have almost zero absolute percent relative error, indicating that the DT model has high accuracy in forecasting the ER for the training, validation, testing, and whole datasets.

3.3.4. Trend analysis (TA) results

Figs. 9–11 display the TA for the D, V_{SG}, and V_{SL}. The effects of the D, V_{SG}, and V_{SL} on the ER are represented. As demonstrated in Fig. 9, increasing the D decreases the ER and shows the proper relations between the D and ER. McLaury et al. [49] confirmed that improving the D reduced the ER. Two points of the D are shown in Fig. 9 because only two measured points have the same V_{SG} and V_{SL}. Therefore, the D TA was studied for only two points and compared with the measured values. Fig. 10 presents a V_{SG} TA, in which the ER is increased by increasing the V_{SG}. Growth in gas velocity exacerbates the ER [50]. In addition, the measured values showed that increasing V_{SG} increases the ER, Fig. 10. The TA of the V_{SL} is shown in Fig. 11. With an increase in the V_{SL}, there is an increase in the ER, indicating a proper relationship to conform to the appropriate PB [51]. revealed that the rising liquid velocity increased the ER. Furthermore, the measured values displayed that the ER is increased by increasing the V_{SL}, Fig. 11.

Consequently, the TA study implies the correct relations between the independent variables, D, V_{SG}, V_{SL}, and the ER, to prove that the DT model conforms to the proper PB.

4. Conclusion

Different machine learning methods are used to determine the erosion rate (ER) based on D, V_{SG}, and V_{SL} using published datasets that were gathered from different references to use wide ranges of data. The various machine learning methods were evaluated and compared to select the best model to determine the ER. The optimum model is the decision tree (DT). The proposed DT model was assessed using various methods. To overcome overfitting or underfitting issues, the gathered datasets were split into three subdivisions, training, validation, and testing datasets. The entire dataset was computed using the proposed model to determine its accuracy. Trend

analyses were performed to confirm the correct PB. Different SEA, namely RMSE and R, were performed to determine the accuracy of the proposed DT model for predicting the ER. The proposed DT model predicted the ER accurately with R of 0.9975, 0.9911, 0.9761, and 0.9908 and AAPRE of 5.0%, 6.27%, 6.26%, and 5.5% for the training, validation, testing, and whole data. The closest leading indicators of the DT model, R, and AAPRE, for training, validation, testing, and entire datasets, indicating the absence of under-fitting or over-fitting problems, demonstrating the robustness of the proposed model. The cross-plotting of training, validation, testing, and whole data also explains the high accuracy of the DT model. The relations between all independent variables (i.e., D, V_{SG}, and V_{SL}) and the dependent variable (i.e., the ER) were demonstrated in this research using the TA. The high accuracy of the proposed model was shown by drawing histograms of absolute percent relative error for the training, validation, testing, and whole datasets.

5. Limitations

The proposed DT model was built based on three independent variables: D, V_{SG}, and V_{SL} with their ranges D of (0.0762–0.1016) m, V_{SG} of (11–41.5) m/s, V_{SL} of (0–0.1630) m/s; however, this model has the high accuracy in determining the erosion rate (ER) of the sand.

Author contribution statement

Fahd Saeed Alakbari: Conceived and designed the analysis; Analyzed and interpreted the data; Contributed analysis tools or data; Wrote the paper.

Mysara Eissa Mohyaldinn, Mohammed Abdalla Ayoub: Conceived and designed the analysis; Analyzed and interpreted the data.

Abdullah Abduljabbar Salih, Azza Hashim Abbas: Conceived and designed the analysis.

Data availability statement

Data will be made available on request.

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.

Acknowledgments

The authors sincerely thank the Yayasan Universiti Teknologi PETRONAS (YUTP FRG Grant No: 015LC0-428) at Universiti Teknologi PETRONAS for supporting this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e17639.

References

[1] A. Abduljabbar, M.E. Mohyaldinn, O. Younis, A. Alghurabi, F.S. Alakbari, Erosion of Sand Screens: A Review of Erosion Prediction Modelling Approaches, Powder Technol., 2022, 117628.
[2] K.G. Jordan, Erosion in Multiphase Production of Oil & Gas, 1998.
[3] S.A. Shirazi, J.R. Shadley, B.S. McLaury, E.F. Rybicki, A Procedure to Predict Solid Particle Erosion in Elbows and Tees, 1995.
[4] R.E. Vieira, A. Mansouri, B.S. McLaury, S.A. Shirazi, Experimental and computational study of erosion in elbows due to sand particles in air flow, Powder Technol. 288 (2016) 339–353.
[5] M.E. Mohyaldinn, M.C. Ismail, M. Ayoub, S.M. Mahmood, Examination and improvement of Salama model for calculation of sand erosion in elbows, in: ICIPEG 2016, Springer, 2017, pp. 687–696.
[6] M.M. Salama, E.S. Venkatesh, Evaluation of API RP 14E Erosional Velocity Limitations for Offshore Gas Wells, 1983.
[7] P. Zahedi, S. Parvandeh, A. Asgharpour, B.S. McLaury, S.A. Shirazi, B.A. McKinney, Random forest regression prediction of solid particle Erosion in elbows, Powder Technol. 338 (2018) 983–992.
[8] R. Kang, H. Liu, A probability model of predicting the sand erosion in elbows for annular flow, Wear 422 (2019) 167–179.
[9] M.E. Mohyaldinn, M.C. Ismail, N. Hasan, A correlation to predict erosion due to sand entrainment in viscous oils flow through elbows, in: Advances in Material Sciences and Engineering, Springer, 2020, pp. 287–297.
[10] Y. Zhang, X. Xu, Solid particle erosion rate predictions through LSBoost, Powder Technol. 388 (2021) 517–525.
[11] S.S. Bahrainian, M. Bakhshesh, E. Hajidavalloo, M. Parsi, A novel approach for solid particle erosion prediction based on Gaussian Process Regression, Wear 466 (2021), 203549.
[12] Y.-Y. Song, L.U. Ying, Decision tree methods: applications for classification and prediction, Shanghai Arch. psychiatry 27 (2) (2015) 130.
[13] M. Almashan, Y. Narusue, H. Morikawa, A Decision Tree Regression Modeling Scheme for Estimating the PVT Properties of Kuwaiti Crude Oil Systems Using Incomplete Datasets, 2019.
[14] C. Carpenter, Decision tree regressions estimate liquid holdup in two-phase gas/liquid flows, J. Petrol. Technol. 73 (11) (2021) 75–76.

[15] J.S. Hernandez, C. Valencia, N. Ratkovich, C.F. Torres, F. Muñoz, Data driven methodology for model selection in flow pattern prediction, Heliyon 5 (11) (2019), e02718.
[16] O.J. Rotimi, et al., Sequential Prediction of Drilling Fluid Loss Using Support Vector Machine and Decision Tree Methods, 2021.
[17] H. Gamal, A. Alsaihati, S. Elkatatny, Predicting the rock sonic logs while drilling by random forest and decision tree-based algorithms, J. Energy Resour. Technol. 144 (4) (2022).
[18] R.E. Vieira, Sand Erosion Model Improvement for Elbows in Gas Production, Multiphase Annular and Low-Liquid Flow, The University of Tulsa, 2014.
[19] M. Parsi, Sand Particle Erosion in Vertical Slug/churn Flow, The University of Tulsa, 2015.
[20] Q.H. Mazumder, Development and Validation of a Mechanistic Model to Predict Erosion in Single-phase and Multiphase Flow, The University of Tulsa, 2004.
[21] M.M. Salama, An Alternative to API 14E Erosional Velocity Limits for Sand Laden Fluids, 1998.
[22] A.T. Bourgoyne, Experimental Study of Erosion in Diverter Systems Due to Sand Production, 1989.
[23] A. Mansouri, H. Arabnejad, S.A. Shirazi, B.S. McLaury, A combined CFD/experimental methodology for erosion prediction, Wear 332 (2015) 1090–1097.
[24] N.R. Kesana, S.A. Grubb, B.S. McLaury, S.A. Shirazi, Ultrasonic measurement of multiphase flow erosion patterns in a standard elbow, J. Energy Resour. Technol. 135 (3) (2013).
[25] P. Zahedi, M. Parsi, A. Asgharpour, B.S. McLaury, S.A. Shirazi, Experimental investigation of sand particle erosion in a 90 elbow in annular two-phase flows, Wear 438 (2019), 203048.
[26] R.E. Vieira, M. Parsi, P. Zahedi, B.S. McLaury, S.A. Shirazi, Ultrasonic measurements of sand particle erosion under upward multiphase annular flow conditions in a vertical-horizontal bend, Int. J. Multiphas. Flow 93 (2017) 48–62.
[27] V. Gudivada, A. Apon, J. Ding, Data quality considerations for big data and machine learning: going beyond data cleaning and transformations, Int. J. Adv. Softw. 10 (1) (2017) 1–20.
[28] F.S. Alakbari, M.E. Mohyaldinn, M.A. Ayoub, A.S. Muhsan, Deep learning approach for robust prediction of reservoir bubble point pressure, ACS Omega 6 (33) (2021) 21499–21513, https://doi.org/10.1021/acsomega.1c02376.
[29] J.W. Tukey, Exploratory Data Analysis, vol. 2, Mass., Reading, 1977.
[30] F.S. Alakbari, M.E. Mohyaldinn, M.A. Ayoub, A.S. Muhsan, I.A. Hussein, A robust Gaussian process regression-based model for the determination of static Young's modulus for sandstone rocks, Neural Comput. Appl. (2023) 1–15.
[31] M. Tom, Mitchell: machine learning, Burr Ridge 45 (37) (1997) 870–877, 1997.
[32] A. Hemmat-Sarapardeh, A. Larestani, N.A. Menad, S. Hajirezaie, Applications of Artificial Intelligence Techniques in the Petroleum Industry, Gulf Professional Publishing, 2020.
[33] J. Quilan, C4. 5, Programs for Machine Learning Morgan Kaufmann, San Mateo, CA, 1993.
[34] T. Amraee, S. Ranjbar, Transient instability prediction using decision tree technique, IEEE Trans. Power Syst. 28 (3) (2013) 3028–3037.
[35] C.J. Sims, L. Meyn, R. Caruana, R.B. Rao, T. Mitchell, M. Krohn, Predicting cesarean delivery with decision tree models, Am. J. Obstet. Gynecol. 183 (5) (2000) 1198–1206.
[36] X.-B. Li, A scalable decision tree system and its application in pattern recognition and intrusion detection, Decis. Support Syst. 41 (1) (2005) 112–130.
[37] A. Priyam, G.R. Abhijeeta, A. Rathee, S. Srivastava, Comparative analysis of decision tree classification algorithms, Int. J. Curr. Eng. Technol. 3 (2) (2013) 334–337.
[38] M. Mathuria, Decision tree analysis on j48 algorithm for data mining, Intrenational J. ofAdvanced Res. Comput. Sci. Soft-ware Eng. 3 (6) (2013).
[39] D.J. Pannell, Sensitivity analysis of normative economic models: theoretical framework and practical strategies, Agric. Econ. 16 (2) (May 1997) 139–152, https://doi.org/10.1111/j.1574-0862.1997.tb00449.x.
[40] M.C. Hill, C.R. Tiedeman, Effective Groundwater Model Calibration: with Analysis of Data, Sensitivities, Predictions, and Uncertainty, John Wiley & Sons, 2006.
[41] A.A. Al-Shammasi, Bubble Point Pressure and Oil Formation Volume Factor Correlations," *Middle East Oil Show and Conference*, vol. 17, Society of Petroleum Engineers, Bahrain, 1999, https://doi.org/10.2118/53185-MS.
[42] S.A. Osman, M.A. Ayoub, M.A. Aggour, Artificial Neural Network Model for Predicting Bottomhole Flowing Pressure in Vertical Multiphase Flow, 2005.
[43] M.A. Ayoub, S.N. Zainal, M.E. Elhaj, K. Ishak, K.E. Hani, Q. Ahmed, Revisiting the Coefficient of Isothermal Oil Compressibility below Bubble Point Pressure and Formulation of a New Model Using Adaptive Neuro-Fuzzy Inference System Technique, 2020.
[44] H.A. Toosi, C. Del Pero, F. Leonforte, M. Lavagna, N. Aste, Machine learning for performance prediction in smart buildings: photovoltaic self-consumption and life cycle cost optimization, Appl. Energy 334 (2023), 120648.
[45] W.-Y. Loh, Regression tress with unbiased variable selection and interaction detection, Stat. Sin. (2002) 361–386.
[46] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Routledge, 2017.
[47] W.-Y. Loh, Y.-S. Shih, Split selection methods for classification trees, Stat. Sin. (1997) 815–840.
[48] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2) (2012).
[49] B.S. McLaury, E.F. Rybicki, S.A. Shirazi, J.R. Shadley, How Operating and Environmental Conditions Affect Erosion, 1999.
[50] W. Qili, J. Binbin, Y. Mingquan, H. Min, L. Xiaochuan, S. Komarneni, Numerical simulation of the flow and erosion behavior of exhaust gas and particles in polysilicon reduction furnace, Sci. Rep. 10 (1) (2020) 1–12.
[51] L. Wu, et al., Experimental study on erosion resistance evaluation of single-layer metal mesh screen, Meas. Control (2021), 00202940211016075.