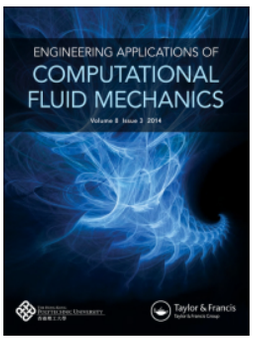

Taylor & Francis
Taylor & Francis Group

Engineering Applications of Computational Fluid Mechanics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tcfm20

Machine learning model development for predicting aeration efficiency through Parshall flume

Sangeeta, Seyed Babak Haji Seyed Asadollah, Ahmad Sharafati, Parveen Sihag, Nadhir Al-Ansari & Kwok-Wing Chau

To cite this article: Sangeeta, Seyed Babak Haji Seyed Asadollah, Ahmad Sharafati, Parveen Sihag, Nadhir Al-Ansari & Kwok-Wing Chau (2021) Machine learning model development for predicting aeration efficiency through Parshall flume, Engineering Applications of Computational Fluid Mechanics, 15:1, 889-901, DOI: 10.1080/19942060.2021.1922314

To link to this article: https://doi.org/10.1080/19942060.2021.1922314

Published online: 17 May 2021.

ENGINEERING APPLICATIONS OF COMPUTATIONAL FLUID MECHANICS
2021, VOL. 15, NO. 1, 889–901
https://doi.org/10.1080/19942060.2021.1922314

Taylor & Francis
Taylor & Francis Group

Machine learning model development for predicting aeration efficiency through Parshall flume

Sangeeta [a], Seyed Babak Haji Seyed Asadollah[b], Ahmad Sharafati [b], Parveen Sihag[c], Nadhir Al-Ansari[d] and Kwok-Wing Chau[e]

[a]Water technology center, Indian Agricultural Research Institute, New Delhi, India; [b]Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran; [c]Civil Engineering Department, Shoolini University, Solan, India; [d]Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, Lulea, Sweden; [e]Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong, People's Republic of China

ABSTRACT

This study compares several advanced machine learning models to obtain the most accurate method for predicting the aeration efficiency (E_{20}) through the Parshall flume. The required dataset is obtained from the laboratory tests using different flumes fabricated in National Institute Technology Kurukshetra, India. Besides, the potential of K Nearest Neighbor (KNN), Random Forest Regression (RFR), and Decision Tree Regression (DTR) models are evaluated to predict the aeration efficiency. In this way, several input combinations (e.g. M1-M15) are provided using the laboratory parameters (e.g. W/L, S/L, Fr, and Re). Different predictive models are obtained based on those input combinations and machine learning models proposed in the present study. The predictive models are assessed based on several performance metrics and visual indicators. Results show that the KNN-M11 model ($RMSE_{testing} = 0.002$, $R^2_{testing} = 0.929$), which includes W/L, S/L, and Fr as predictive variables outperforms the other predictive models. Furthermore, an enhancement is observed in KNN model estimation accuracy compared to the previously developed empirical models. In general, the predictive model dominated in the present study provides adequate performance in predicting the aeration efficiency in the Parshall flume.

ARTICLE HISTORY
Received 21 August 2020
Accepted 22 April 2021

KEYWORDS
Aeration efficiency; Parshall flume; prediction; machine learning models

1. Introduction

Parshall flume is one of the most commonly used fixed hydraulic structures to measure the surface water and irrigation flows. For a long time, numerous investigators focused on one or more flume components to simplify, improve, and refine its design and operation. Result of those intense experimental researches, the Parshall flume was eventually developed. A brief review of Parshall flume development is offered to benefit those unfamiliar with its history. Cone (1917) initiated a flume made up of a converging section, a diverging section, and a short throat section in-between them. His 'venturi flume' had the ground part as flat throughout the length. Parshall and Rohwer (1921) and Parshall (1928) had designed the present flume and expanded its applications in agriculture. His designed Parshall flume has an upstream converging section, end with downstream diverging exit section & in-between flat throat that has defined width with a downward sloped floor & upward slope section. Various shape weirs & Venturi flumes were used to measure flow at that time, but these devices had various limitations and disadvantages. Six years after his work began on the development of 'modified venturi flume', Dr. Parshall filed for the patent of his construction of a new 'Parshall flume'. Then these flumes were installed in various American irrigation facilities (Heiner & Barfuss, 2011).

A Parshall flume includes i) converging, ii) throat, and iii) diverging portions. The converging portion made up of an upstream narrowing approach with a level floor followed by a throat portion with downward-sloping, then flow continues to the diverging downstream portion with upward-sloping as shown in Figure 1 (Mustonen, 1986; Parshall, 1950).

The accelerated velocity causes the air injection process in the Parshall flume (Reclamation et al., 1997). The flow velocity accelerates through contracting sidewalls in the converging portion; then, the flow condition changes from subcritical to supercritical in the throat portion due to contraction and drop. Afterward, aeration is performed in the diverging portion by changes the supercritical flow to subcritical.

CONTACT Ahmad Sharafati asharafati@srbiau.ac.ir, Nadhir Al-Ansari nadhir.alansari@ltu.se

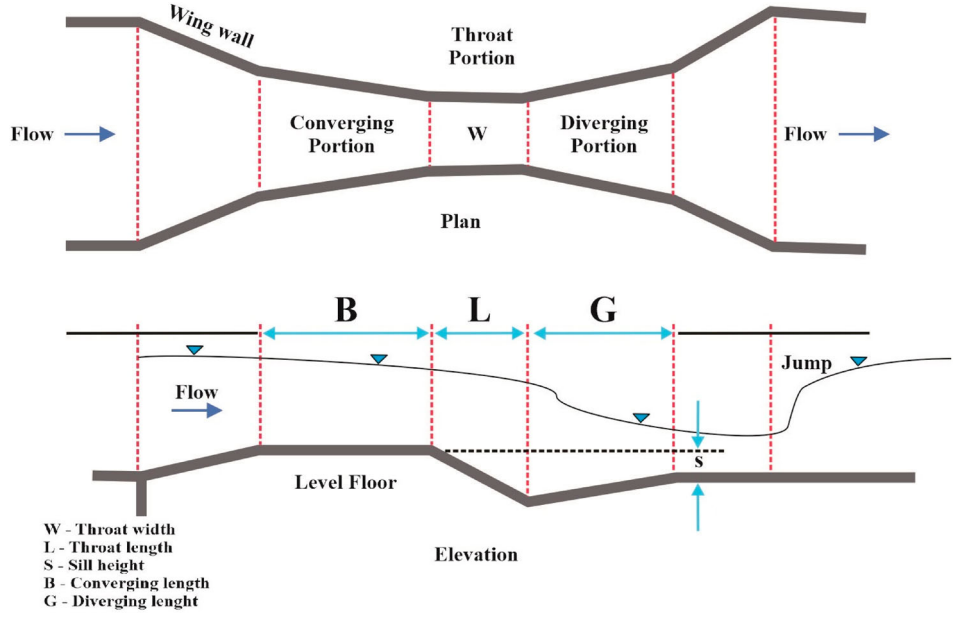

Figure 1. Definition sketch of Parshal flume.

The dissolved oxygen (DO) concentration is an influential factor in biological and chemical activities in aquatic ecosystems, such as the self-purification of the rivers (Asadollah et al., 2020; Sharafati et al., 2020a). A minimum level of DO is necessary for the survival of the aquatic life, and thus several problems would be raised for the aquatic life when the DO concentration drops down to 5 mg/l. Aeration is a common approach to inject air into the water body to increase the DO level. Besides, it is an essential treatment process in wastewater treatment plants.

(Wilhelms et al., 1993), (Gulliver et al., 1998), (Ervine, 1998) and (Chanson, 1995) examined the aeration process in hydraulic structures. (Kaya & Emiroglu, 2010) studied oxygen transfer at baffled chutes. Although weir is a common hydraulic structure to increase DO levels into the treatment channels, it is inappropriate in channels with mild slopes. In contrast, flume provides an efficient aeration system in treatment plants (Dursun, 2016).

The oxygen disperses in the flowing water by turbulent mixing (hydraulic jump) and molecular diffusion (bubble formation). There are two laminar layers on each side of the water–air interface to control the oxygen dispersion. The oxygen transfer rate is directly related to the concentration gradient and would be computed as follows:

$$\frac{dt}{dc} = K_L \frac{A}{V}(C_S - C) \tag{1}$$

where C and C_s are the absorbed and saturated concentration of oxygen, respectively. K_L is the coefficient of the liquid layer, A and V are the area, and the volume of transferred oxygen and t signifies time.

(Gulliver et al., 1998) developed a time-independent formula to compute the oxygen aeration efficiency, E, as follows:

$$E = \frac{C_D - C_U}{C_S - C_U} = 1 - \frac{C_S - C_D}{C_S - C_U} = 1 - \frac{1}{R} \tag{2}$$

where, C_U and $C_D$are the oxygen concentration in upstream and downstream zones, respectively, R represents the oxygen aeration deficient ratio.

The aeration efficiency is unity in the full transfer of oxygen to water, while the zero value indicates that no oxygen can be transferred. To preserve the homogeneity in measured experiments, the results obtained in different temperatures are normalized at $20^{\circ}C$ using the following equation (Gulliver et al., 1998):

$$1 - E_{20} = (1 - E)^{1/f} \tag{3}$$

where, E_{20} is the oxygen transfer efficiency at $20^{\circ}C$ and f, which is the aeration exponent, would be computed as follows:

$$f = 1 + 2.1 * 10^{-2}(T - 20) + 8.25 * 10^{-5}(T - 20)^2 \tag{4}$$

Several conventional empirical formulas were developed by previous authors (Markofsky & Kobus, 1978; Preul & Holler, 1969; Wormleaton & Tsang, 2000). The details of the models are presented in Table 1.

Several investigations have been conducted to assess the aeration efficiency of hydraulic structures. (Dursun, 2016) and (Tiwari & Sihag, 2020) assessed the

Table 1. Conventional formulas proposed in the previous studies to estimate the aeration efficiency.

Scholar	Formula
Preul and Holler (1969)	$E_{20} = 1 - \frac{1}{1 + 666Fr^{-3.33}}$
Avery and Novak (1978)	$E_{20} = 1 + \left[\frac{1}{1 + 0.24 * 10^{-4} Fr^{1.78} R_e^{0.53}}\right]^{1.115}$
Markofsky and Kobus (1978)	$E_{20} = 1 - \left[\frac{1}{1 + 0.1Fr^{1.2}}\right]^{1.115}$
Wormleaton and Tsang (2000)	$E_{20} = 1 - [1 + 0.385 * 10^{-6} Fr^{2.297} R_e^{0.684}]^{-1}$

Note; Fr: Froude number, Re: Reynolds number.

aeration efficiency using the experiments on Parshall flumes. (Emiroglu & Baylar, 2003) performed aeration on a steeped channel either with or without end sills. (Chanson, 2002) calibrated the gas–liquid interface through a stepped channel.

The application of soft computing models has been drastically increased in different engineering fields (Sihag and Vajesnayee, 2018; Kumar et al., 2018, 2020; Sihag et al., 2017, 2019; Tiwari et al., 2019). Very few studies have been conducted to assess the aeration mechanism in flumes (i.e. Parshall and modified venture flumes). (Dursun, 2016) provided an experimental laboratory dataset to assess the aeration of the small. (Tiwari et al., 2019) examined the potential of soft computing models, including adaptive neuro-fuzzy inference system (ANFIS), fuzzy logic (FL), and artificial neural network (ANN), to simulate the transfer efficiency of oxygen through the Parshall flumes.

Machine learning models, which are the newest subset of soft computing, are divided into various subcategories from such supervised, unsupervised, and reinforced learning methods can be highlighted. Supervised learning proved to be more applicable to machine learning, consisting of several algorithms employed on hydraulic problems. As a widely employed supervised method, the Support Vector Machines were used in various prediction tasks related to a diverse range of problems such as optical lenses (Petković et al., 2014b), robotics (Jović et al., 2016) as well as the hydraulic jump oxygen transfer (Tiwari, 2019). The Fuzzy-based algorithms also proved to be a highly efficient forecasting technique which utilized in various scientific problems like underwater laser cutting parameter selection (Nikolić et al., 2016), management of precipitation concentration (Petković et al., 2017), wind turbines (Petković et al., 2014a) and many more (Mohammed et al., 2020; Sharafati et al., 2020b, 2021).

While these methods are considered the machine learning most commonly used algorithms, newer models such as tree-based and extreme learning models have also represented high prediction accuracy. For example, M5P, as a basic tree-based model, has been utilized in the assessment of Parshall flumes aeration efficiency (Ranjan & Tiwari, 2019; Sangeeta & Tiwari, 2019) and proved to be highly efficient against alternative AI algorithms. As another novel algorithm, Extreme Machine Learning also proved to be a useful prediction tool for weirs discharge coefficient problems (Gharib et al., 2020; Yarmohammadi et al., 2019).

Overall, the measurement of Parshall flow aeration efficiency is very costly and time-consuming, which requires laboratory experiments. Due to this matter, the primary motivation of the current study is to evaluate the E20 using novel machine learning algorithms by employing the least experimental parameters and saves time, money, and effort. So, this study is helpful for hydraulic researchers and scientists.

This study evaluates the accuracy of three supervised machine learning approaches, in the particular Decision tree and Random forest as advanced tree-based algorithms and K Nearest Neighbor as an instant based algorithm to predict the oxygen aeration efficiency (E20) in the small Parshall flumes with different forms. To the best of the authors' knowledge, the proposed algorithms have not been developed yet to predict the oxygen aeration efficiency of the Parshall flumes. Furthermore, the results obtained from the best predictive model are compared with the empirical models reported in the literature.

2. Materials and methods

2.1. K-Nearest Neighbor

As a pattern recognition regression technique, K Nearest Neighbors (KNN) usually implemented two general tasks, including density function evaluation and subsequently classification of test dataset due to their distribution and pattern. The initial phase of the algorithm consists of finding an approach to measure the distance related to training and testing data. The Euclidean method is the most common technique for distance determination between the mentioned two sets of data:

$$d(\alpha, \beta) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

Where α and β denote the training data $[x_1, \ldots, x_n]$ and their corresponding $[y_1, \ldots, y_n]$ parameters, respectively. The dataset samples were sorted from minimum to maximum based on the assessed Euclidean distance (d), which symbolized the most and least similarity, respectively.

Table 2. Values of parameter used in KNN, RFR and DTR algorithms.

KNN		RFR		DTR	
Parameters	Value	Parameters	Value	Parameters	Value
N_neighbors	5	N_estimators	10	Splitter	Random
Weights	uniform	Max_depth	100	Max depth	5
Algorithm	brute	Min samples split	20	Min samples split	30
Leaf size	30	Random state	0	Random state	3

In the next stage, the K,, an essential parameter in evaluating a specific dataset features, must be determined. Measuring the number of neighbors (K) is crucial, and the effectiveness of this regression method depends on choosing the most similar samples from the original dataset. A small K results in single-point sensitivity, while large-scale K outcomes overlap samples from other clusters into the desired neighborhood. It is evident from previous literature that the optimal value of K assessed by employing the cross-validation technique (Shabani et al., 2020; Wu et al., 2008).

2.2. Decision tree

A decision tree, a supervised algorithm in machine learning, can undertake both regression and classification tasks. The Decision Tree Regression (DTR) technique extracts major features from a database and structure them in a tree-shaped architecture, where internal and terminal nodes, respectively, exhibit splits and leaves. By gathering individual trees, which have their specific rules, a collection of rules is starting to form that would be utilized in the regression stage (Witten & Frank, 2002).

Initially, tree bough is established using the original training data. Then the DTR algorithm creates branches from the main body of the tree and subsequently new branches from the old one by separating the data using a binary split procedure. The advance of this process stops when a branch becomes too small and impossible to segregate, which in that case, the related node constructs the leave of a tree or the terminal node (Kamiński et al., 2018; Quinlan, 1987).

What makes the DTR an applicable machine learning method is the easy interpretation of rules originated from the trees, which usually exhibit a logical pattern instead of mathematical relations. Though, the inability to handle nonlinear and noisy data and time series problems are considered major DTR drawbacks (Mitchell, 1997; Tso & Yau, 2007).

2.3. Random forest

Random forest is an ensemble algorithm that originated from tree practice, widely used in machine learning and artificial intelligence (Liaw & Wiener, 2002). It is mostly used in a classification problem, such as regression (Goel et al., 2017). The random forest regression (RFR) function is to grow several decisions or predicting trees during the learning process. All of these trees are trained individually. Finally, the result is obtained by identifying the specified category based on classified states and/or the averaged prediction of individual trees (Barandiaran, 1998; Ho, 1995).

The random forest learning algorithm employs generic bootstrap and bagging techniques to train the tree. Bootstrap enhances the model performance by reducing the variance and without introducing any changes in the bias.

Breiman (2001) considered RF as a set of n stochastic decision trees $\{DT(x, \Theta_i, l), i = 1, \ldots, n\}$, where $DT(x, \Theta_i, l)$ indicates the i-th predicting stochastic tree or $DT(x)$, which is grown by $\{\Theta_i\}$ as a uniform independent stochastic distribution vector. This vector encrypts the stochastic arrangement needed for constructing the tree, and l is the growth learning data. $\{\Theta_i\}$ is selected prior to the tree growth and is independence from learning data. All the predicting trees are integrated and averaged so that an estimator forest of $DT(x)$ is constructed.

$$DT(x.\Theta_1, \ldots, \Theta_i.l) = \frac{1}{K}\sum_{i=1}^{n} DT(X.\Theta_i.l) \qquad (6)$$

The algorithms evaluated in the current study are developed utilizing the Scikit learn library on Python programming language over the Anaconda platform. While there are numerous effective parameters in the KNN, DTR, and RFR algorithms, in the current study, only the parameters listed in Table 2 proved to be highly influential on final results, and the other variables have been set to default.

2.4. Description of laboratory experiments

The oxygen transfer efficiency of small Parshal flumes was obtained experimentally in hydraulics at the National Institute Technology in Kurukshetra, India (Sangeeta, 2018). Several flumes were fabricated at the workshop of the mechanical department of the same institute. The dimensional details of the flumes are presented in Table 3.

The models are firmly fixed in tilting rectangular rigid steel channel of $0.25 \times 0.30 \times 4.00m$ as width, height,

Table 3. Description of test-models dimension.

Experimental models	W (cm)	L (cm)	S (cm)
EM-1	2.54	7.62	1.91
EM-2	2.54	7.62	3.91
EM-3	2.54	7.62	5.91
EM-4	2.54	10	1.91
EM-5	2.54	10	3.91
EM-6	2.54	10	5.91
EM-7	5.08	11.45	2.22
EM-8	5.08	11.45	4.22
EM-9	5.08	11.45	6.22
EM-10	5.08	15	2.22
EM-11	5.08	15	4.22
EM-12	5.08	15	6.22

Note; W: Throat width, L: Throat length, S: Sill height.

and length dimensions, respectively. The tilting channel has a transparent acrylic sheet in the middle for 1.8 m in length. To supply the flow required for experiments, a centrifugal pump is utilized with a discharge capacity of $6\frac{m^3}{s}$ to provide the water recirculation. A headbox is used at the begging of the flume, while the upstream entry of the flume is equipped with a metal screen to damp the water fluctuation. The water discharge is regulated and measured using a regulating valve and orifice-meter, respectively. The measurement procedure of water depth is performed by a digital gauge with an accuracy of 0.01 mm. Figure 2 illustrates the schematic structure of the test setup.

To find out the oxygen aeration performance of the experimental models, the storage cum aeration tank is filled with a fixed volume of water through each experiment. The sodium sulphite (Na2SO3) and cobalt chloride (CoCl2) are thoroughly mixed in the aeration tank to reduce DO level in the range of 1–2 mg/l. Several samples of water are collected to measure the initial DO concentration. Then, the other samples are obtained after 1-minute of running the flume. It is essential to preserve the level of DO less saturation condition. The Azide modification method is used to compute the level of DO. Besides, a mercury thermometer is employed to measure the temperature of water in the tank. Then, the oxygen aeration efficiency, E20, is computed using Eqs.2–4. A similar procedure is conducted for each experimental model. Ultimately, 237 test runs are obtained while the doubtful tests are repeated twice to achieve reliable results.

2.5. Description of the proposed predictive models

The combinations of non-dimensional predictors (e.g. $W/L, S/L, Fr,$ and Re) are defined to obtain the best predictive model for simulating the aeration efficiency in the Parshall flume. In this way, fifteen combinations ($M1 - M15$)are defined to assess supervised machine learning models (e.g. *DTR*, *KNN* and *RFR*). The details of input combinations defined in the present study are illustrated in Table 4.

2.6. Description of dataset

Several non-dimensional physical and hydraulic parameters, including $W/L, S/L, Fr, and Re$ are measured to provide a dataset for simulating the aeration efficiency in Parshal flume ($E20$). The collected experimental dataset

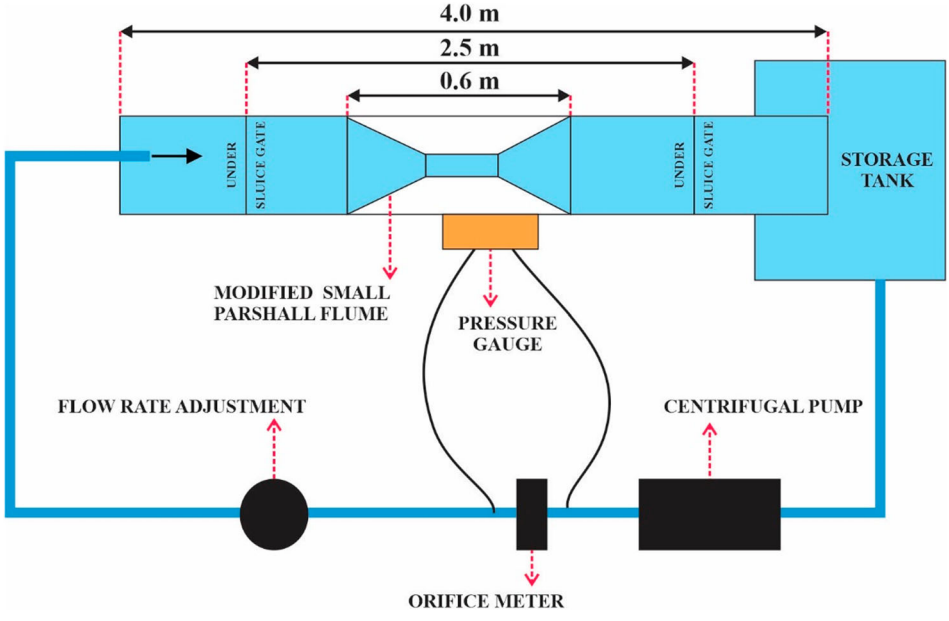

Figure 2. A schematic view of test set up.

Table 4. Description of different proposed input variable combinations to predict the aeration efficiency.

Input combinations models	Predictive variables			
	W/L	*S/L*	*Fr*	Re
M1	✓			
M2		✓		
M3			✓	
M4				✓
M5	✓	✓		
M6	✓		✓	
M7	✓			✓
M8		✓	✓	
M9		✓		✓
M10			✓	✓
M11	✓	✓	✓	
M12	✓	✓		✓
M13	✓		✓	✓
M14		✓	✓	✓
M15	✓	✓	✓	✓

is split into the training and testing subsets to develop the soft computing models—the training subset employed for model preparation. In contrast, the testing dataset is used for model validation. The dataset is divided into 163–74 observations for the training-testing stages. The division is obtained by trial and error search to provide the best performance. Table 5 demonstrates the ranges of measured target and input variables over both training and testing stages.

2.7. Description of performance metrics

To analyze the capability of various modeling methods to predict the E_{20}, coefficient of determination (R^2), root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE), and mean absolute error (MAE) indices are calculated in both training and the testing stages as following equations:

$$R^2 = \left[\frac{n\sum E_{obs}E_{pred} - (\sum E_{obs})(\sum E_{pred})}{\sqrt{n(\sum {E_{obs}}^2) - (\sum E_{obs})^2}\sqrt{n(\sum {E_{pred}}^2) - (\sum E_{pred})^2}} \right]^2 \quad (7)$$

$$MAE = \frac{1}{n}|E_{obs} - E_{pred}| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n}(\sum_{i=1}^{n}(E_{obs} - E_{pred})^2} \quad (9)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(E_{obs} - E_{pred})^2}{\sum_{i=1}^{n}(E_{obs} - \overline{E_{obs}})^2} \quad (10)$$

Where, E_{obs} and E_{pred} are observed and predicted aeration efficiency, respectively. n is the total number of observations.

3. Results and discussions:

Several predictive models with various input combinations are developed to achieve the best predictive model for predicting the aeration efficiency (E20) at Parshall flumes. In this way, fifteen possible input combinations are obtained based on the four input variables (Table 4). The predictive models proposed in the present study, including Decision Tree, K nearest neighbor, and Random Forest Regression models, use those combinations to predict the E_{20} at Parshall flumes.

Tables 6–8, respectively, report the performance indices obtained for different combinations of DTR, KNN, and RFR models in both training and testing stages. From Table 6, it is clear that the DRT- M15 model ($RMSE_{test}$ = 0.003 and NSE_{test} = 0.867) provides better performance compared to other input combinations.

Table 5. Statistical characteristics of the datasets used in training and testing stages.

Statistical characteristic	Stage	*W/L*	*S/L*	*Fr*	*Re*	*E20*
Minimum	Training	0.254	0.148	0.109	3139.38	0.033
	Testing	0.254	0.148	0.109	3139.38	0.033
Maximum	Training	0.444	0.776	0.175	41573.13	0.093
	Testing	0.444	0.776	0.175	41573.13	0.093
Mean	Training	0.342	0.384	0.141	20065.633	0.061
	Testing	0.344	0.397	0.142	19734.552	0.061
Median	Training	0.333	0.369	0.117	20786.560	0.060
	Testing	0.339	0.391	0.141	18836.290	0.059
Standard Deviation	Training	0.068	0.183	0.028	10767.624	0.012
	Testing	0.068	0.185	0.028	11549.004	0.013
Kurtosis	Training	−1.007	−0.589	−1.931	−0.893	−0.177
	Testing	−0.993	−0.550	−1.918	−1.043	−0.052
Skewness	Training	0.293	0.563	0.145	0.188	0.421
	Testing	0.269	0.528	0.122	0.291	0.394

Table 6. Performance indices parameters for different input combination based DTR models.

	Training stage				Testing stage			
Input Combination	MAE	RMSE	R2	NSE	MAE	RMSE	R2	NSE
M1	0.004	0.006	0.769	0.769	0.005	0.006	0.768	0.767
M2	0.006	0.007	0.669	0.669	0.006	0.007	0.671	0.670
M3	0.008	0.009	0.454	0.454	0.008	0.009	0.460	0.458
M4	0.009	0.011	0.174	0.174	0.009	0.012	0.065	0.057
M5	0.005	0.006	0.744	0.744	0.005	0.006	0.749	0.749
M6	0.004	0.005	0.848	0.848	0.004	0.005	0.836	0.834
M7	0.010	0.012	0.024	0.024	0.010	0.012	0.007	0.001
M8	0.007	0.008	0.534	0.534	0.007	0.009	0.528	0.527
M9	0.008	0.010	0.353	0.353	0.008	0.010	0.397	0.395
M10	0.008	0.009	0.466	0.466	0.008	0.010	0.389	0.388
M11	0.004	0.005	0.840	0.840	0.004	0.005	0.833	0.833
M12	0.006	0.008	0.560	0.560	0.005	0.007	0.657	0.650
M13	0.004	0.005	0.806	0.806	0.005	0.006	0.745	0.741
M14	0.005	0.006	0.725	0.725	0.005	0.006	0.732	0.732
M15	**0.003**	**0.004**	**0.878**	**0.878**	**0.003**	**0.005**	**0.868**	**0.867**

Table 7. Performance indices parameters for different input combination based KNN models.

	Training Phase				Testing Phase			
Input Combination	MAE	RMSE	R2	NSE	MAE	RMSE	R2	NSE
M1	0.004	0.006	0.750	0.741	0.005	0.006	0.740	0.735
M2	0.004	0.006	0.771	0.769	0.004	0.006	0.754	0.753
M3	0.008	0.010	0.398	0.380	0.008	0.011	0.300	0.284
M4	0.008	0.009	0.430	0.422	0.008	0.010	0.323	0.295
M5	0.004	0.006	0.793	0.789	0.004	0.006	0.787	0.783
M6	0.003	0.004	0.869	0.869	0.004	0.005	0.818	0.818
M7	0.004	0.006	0.799	0.795	0.006	0.008	0.576	0.563
M8	0.002	0.003	0.932	0.931	0.003	0.004	0.916	0.915
M9	0.008	0.009	0.524	0.523	0.008	0.009	0.482	0.470
M10	0.008	0.009	0.418	0.410	0.008	0.010	0.323	0.294
M11	**0.002**	**0.003**	**0.936**	**0.935**	**0.002**	**0.003**	**0.929**	**0.928**
M12	0.007	0.008	0.570	0.566	0.008	0.009	0.492	0.482
M13	0.004	0.006	0.801	0.797	0.006	0.008	0.590	0.577
M14	0.008	0.009	0.523	0.523	0.008	0.009	0.481	0.470
M15	0.007	0.008	0.572	0.568	0.007	0.009	0.498	0.486

Table 8. Performance indices parameters for different input combination based RFR models.

	Training Phase				Testing Phase			
Input Combination	MAE	RMSE	R2	NSE	MAE	RMSE	R2	NSE
M1	0.004	0.006	0.768	0.768	0.005	0.006	0.770	0.768
M2	0.004	0.005	0.818	0.818	0.004	0.005	0.843	0.842
M3	0.008	0.009	0.461	0.461	0.008	0.010	0.409	0.409
M4	0.008	0.010	0.362	0.341	0.009	0.012	0.149	0.147
M5	0.004	0.005	0.820	0.820	0.004	0.005	0.839	0.839
M6	0.003	0.005	0.866	0.866	0.004	0.005	0.822	0.821
M7	0.003	0.004	0.868	0.868	0.004	0.005	0.828	0.827
M8	0.004	0.005	0.849	0.844	0.004	0.005	0.844	0.840
M9	0.004	0.005	0.824	0.814	0.004	0.005	0.867	0.845
M10	0.008	0.009	0.463	0.463	0.008	0.010	0.381	0.378
M11	**0.003**	**0.004**	**0.913**	**0.910**	**0.003**	**0.004**	**0.886**	**0.884**
M12	0.003	0.004	0.886	0.886	0.003	0.005	0.870	0.870
M13	0.003	0.004	0.867	0.867	0.004	0.005	0.825	0.823
M14	0.004	0.005	0.842	0.835	0.004	0.005	0.842	0.833
M15	0.003	0.004	0.912	0.910	0.003	0.004	0.882	0.881

Besides, KNN-M11 ($RMSE_{test}$ = 0.003 and NSE_{test} = 0.938) offers the most accurate prediction among all KNN models (Table 7). Furthermore, better prediction performance is observed in RFR-M11 ($RMSE_{test}$ = 0.004 and NSE_{test} = 0.884) compared with the other RFR-based models (Table 8).

The result obtained from the comparison between the best predictive models (e.g. DRT-M15, KNN-M11, and RFR-M11) indicates that the KNN-M11 provides better performance prediction than other machine learning techniques. In contrast, the KNN-M11 model comprises a simpler input combination including W/L, S/L, and Fr.

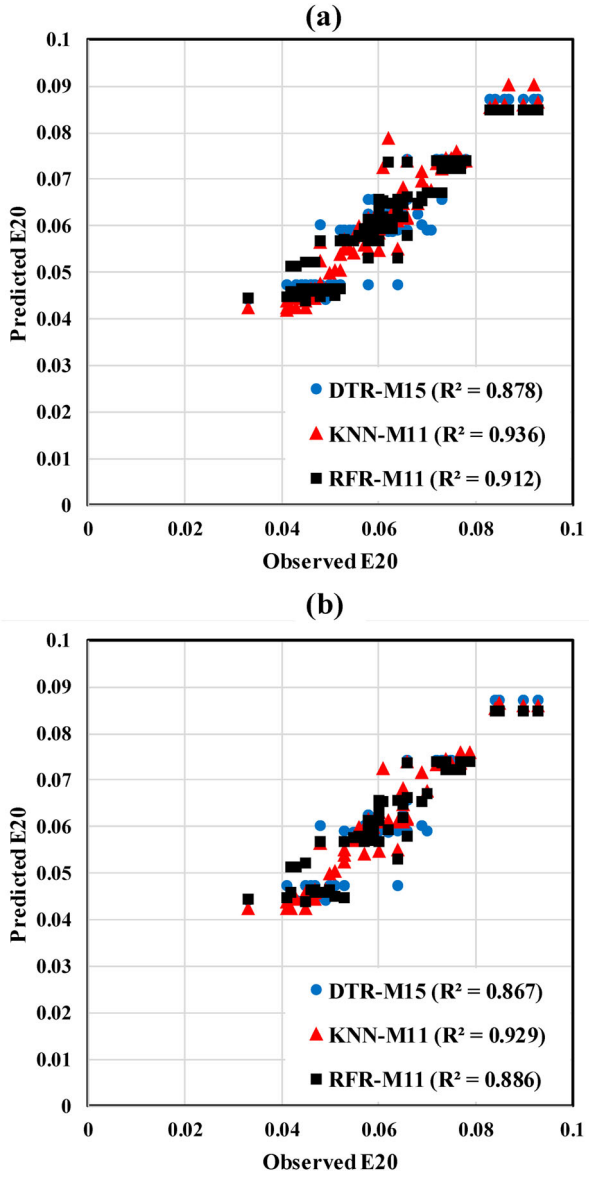

Figure 3. Agreement plot among actual vs predicted values of E_{20} with various soft computing techniques using (a) training dataset (b) testing data set.

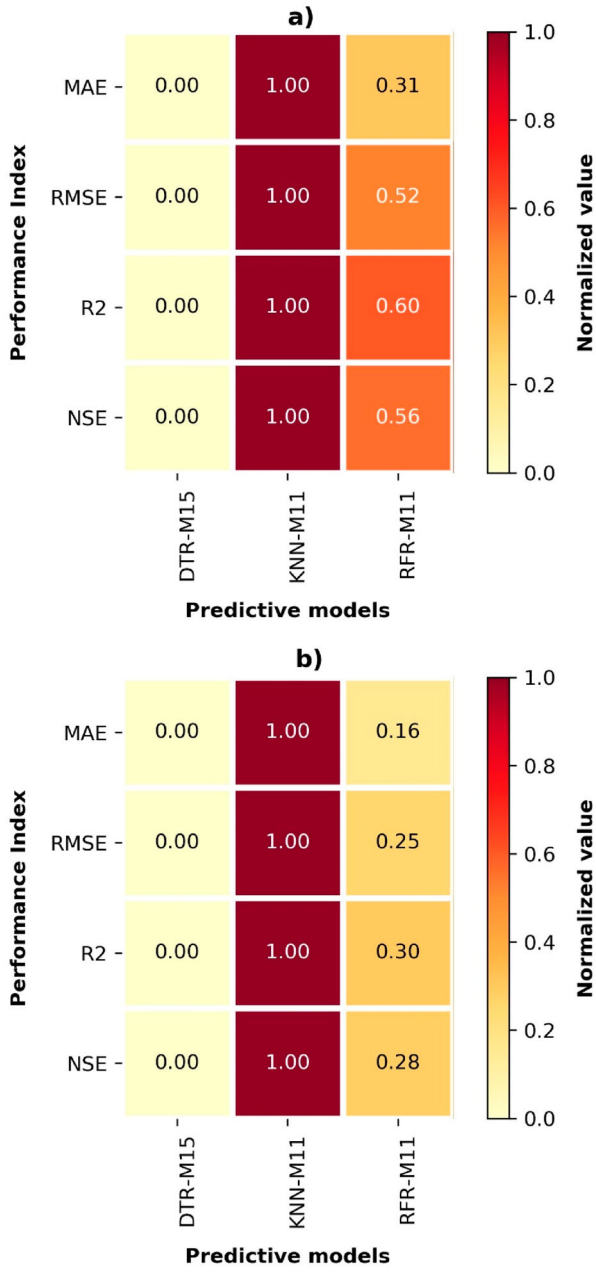

Figure 4. Heat plot of the observed and the predicted values of E_{20} at Parshall flumes. (a) training dataset (b) testing dataset.

The superiority is notified as a significant reduction in RMSE & MAE and considerable improvement in R^2 & NSE obtained by KNN rather than other techniques. Outcomes of training and testing stages suggest that both these stages have the same results, which means the developed models are not data sensitive and have similar behavior with different data samples.

To investigate the linear relation among observed and predicted values of E_{20} at Parshall flumes, the agreement plots of the predictive models are shown in Figure 3. Although all the predictive models provide acceptable performance in predicting the E_{20} at Parshall flumes, the KNN-M11 model has the highest value of the coefficient of determination (R^2) value as 0.929 for the testing stage. It is indicated that the values predicted by KNN-M11 models are lying significantly closer to the line of perfect agreement ($R^2 = 1.00$). This indicates that the KNN learning method is more suitable than other models for predicting E20 at Parshall flumes.

The graphical comparison among predictive models in terms of standardized performance evaluation parameters is also shown in the heat-map plot. As shown in Figure 4, the KNN-M11 model (dark red) offers the best prediction performance compared to other models. In contrast, the lowest performance is found in the

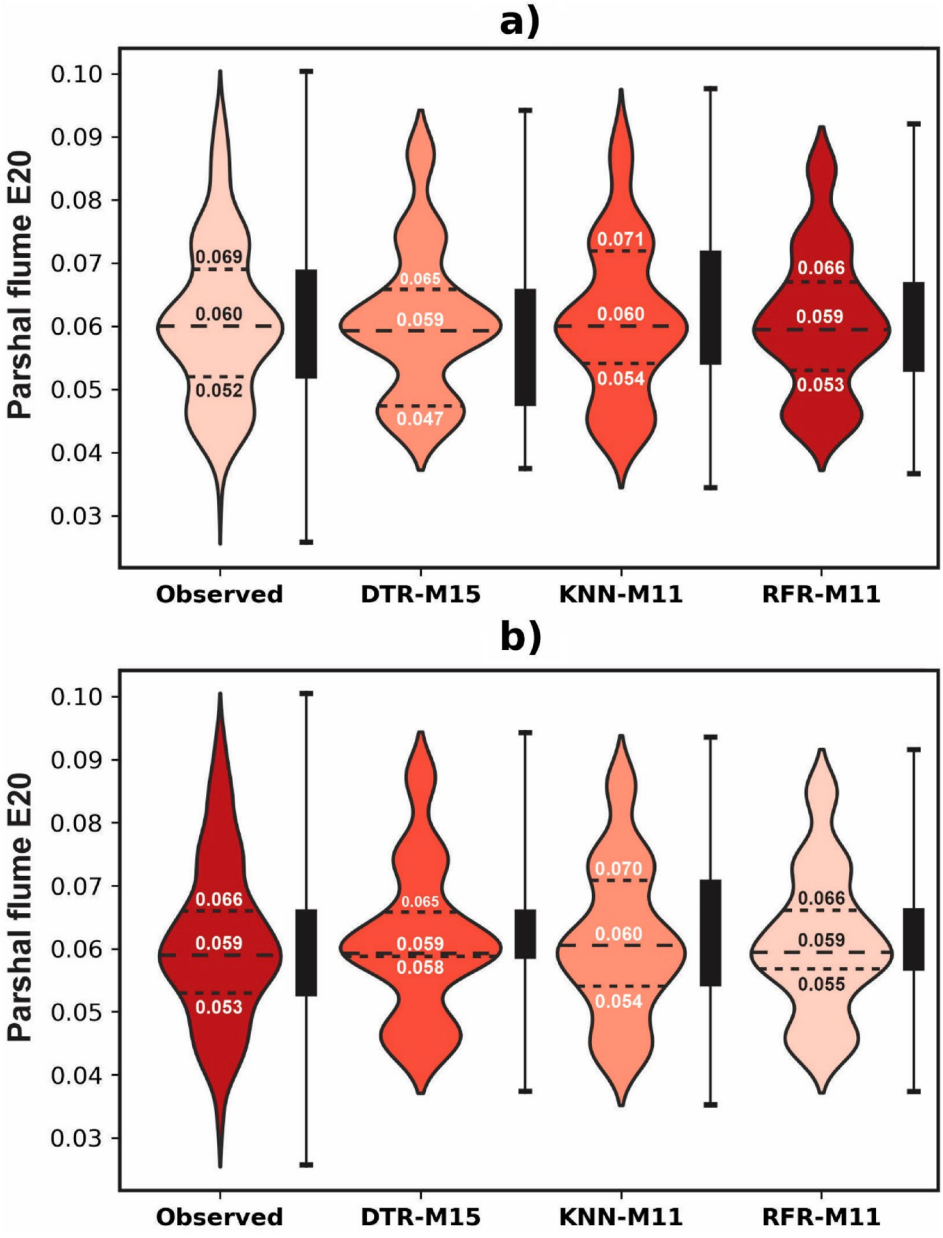

Figure 5. Violin diagram of the observed and the predicted values of E_{20} at Parshall flumes. (a) training dataset (b) testing dataset.

DTR-M15 model (Yellow) to predict the aeration efficiency (E_{20}) at Parshall flumes.

To evaluate the consistency of the predicted E_{20} against the observed values, the 25%, 50%, 75% quantile values of the experimental and predicted E_{20} and their corresponding Inter Quartile Range (IQR) are assessed using a combination of Violin and Box plots. From Figure 5, it is evident that in the training stage, KNN-M11 (IQR = 0.017) profoundly mimics the observed data (IQR = 0.017), in the testing stage, the KNN-M11 (IQR = 0.016) and RFR-M11 (IQR = 0.011) models represent very close prediction to the experimental values (IQR = 0.013), and the DTR-M15 (IQR = 0.007) exhibits the least accurate prediction.

The relative error diagrams of DTR-M15, KNN-M11, and RFR-M11 and their boxplots in the testing stage are depicted in Figure 6. The figure indicates that the lowest confidence limit of the relative error is observed in KNN-M11 in the range of −8.67% to 12.12%. In contrast, the RFR-M11 provides the biggest confidence limit in the range of −17.16% to 18.32%.

As one of the most commonly utilized graphical evaluation approaches, the Taylor diagram is represented in Figure 7. This diagram combines the statistical measurement with the correlation coefficient and assigns a colorful shape to predictive models. The performance of models then assessed by considering their distance with experimental values. The figure shows that the KNN-C11

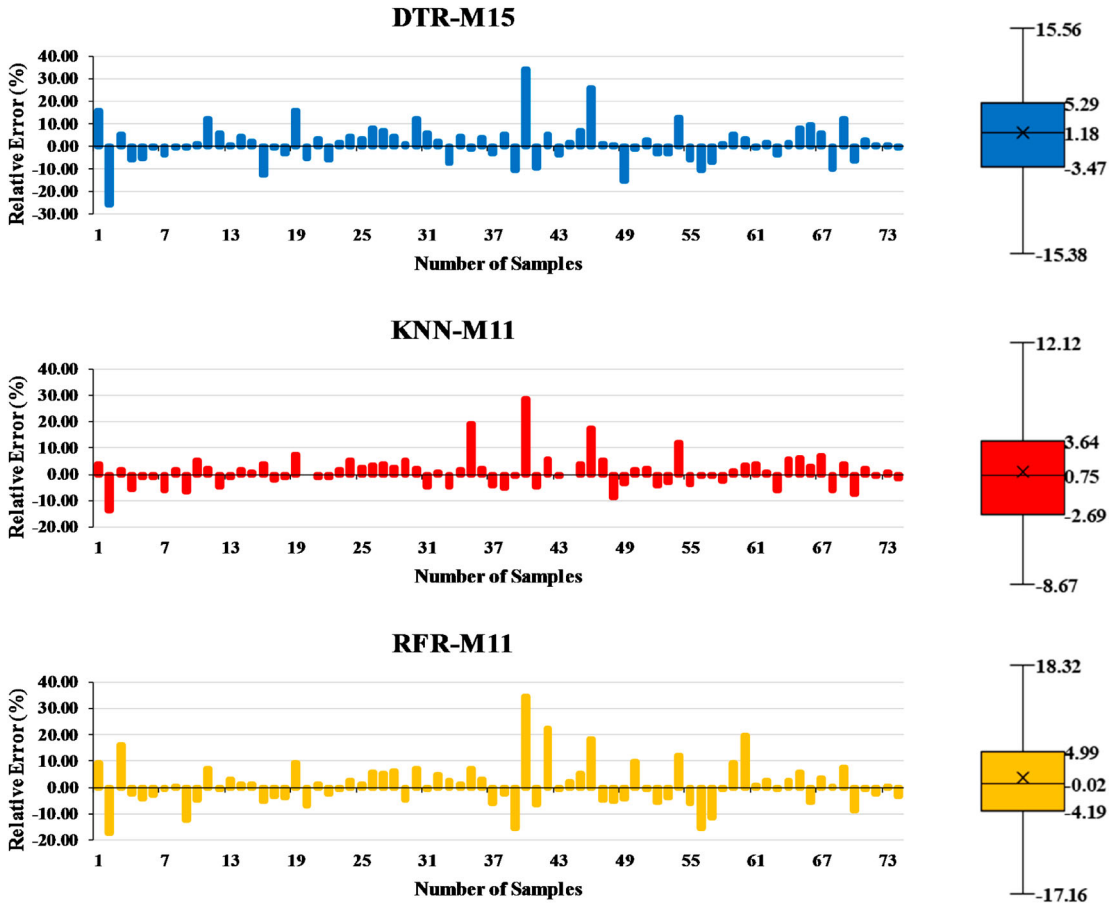

Figure 6. Relative error plot of the predicted values of E_{20} at Parshall flumes.

(orange dot) is the closest shape with the experimental (dark blue dot) and subsequently considered a more accurate model. Meanwhile, with a slightly more distance, the DTR-C15 is ranked as least precise among others.

Overall it can be concluded that, while all three employed machine learning algorithms in the current study provide the acceptable prediction performance, the KNN using the M11 combination, which includes W/L, S/L, and Fr input parameters, shows a slightly higher prediction accuracy compare to Decision Tree and Random forest models based on both statistical and graphical evaluations. It was evident that the KNN-M11 improved the prediction accuracy by reducing the RMSE by nearly 40% as well as enhancing the R^2 by 7% over the testing phase compare to other alternatives.

From Tables 6–8, it is evident that the prediction performance significantly decreases in M4 ($R^2_{testing\ average} = 0.179$) and M3 ($R^2_{testing\ average} = 0.389$) combinations which indicate the low correlation of Re and Fr input parameters to the E_{20} values, respectively. However, the most important input parameters find to be the W/L with average $R^2_{testing}$ of 0.76 compared to other inputs.

Table 9 compares the outcome of the current study and the result obtained from calculating the E20 with the previous experimental formula proposed by (Avery & Novak, 1978; Markofsky & Kobus, 1978; Preul & Holler, 1969; Wormleaton & Tsang, 2000). As presented in Table 8, it is clear that the machine learning algorithm proposed in the present study significantly outperforms the empirical formulas in predicting E20.

4. Conclusions

Prediction of the aeration efficiency (E_{20}) values at Parshall flumes is essential because it has a proper application in irrigation canals, mine discharge, dam seepage, sewage treatment plants. In this study, experimental data are used to investigate the performance of machine learning approaches, namely K Nearest Neighbor (KNN), Random Forest Regression (RFR), and Decision Tree Regression (DTR). Laboratory parameters consist of W/L, S/L, Fr, and Re are employed as inputs of prediction models, and subsequently, by arranging these four parameters, 15 input combinations are established, which are noted as $M1 \sim M15$. A total of 237 experimental

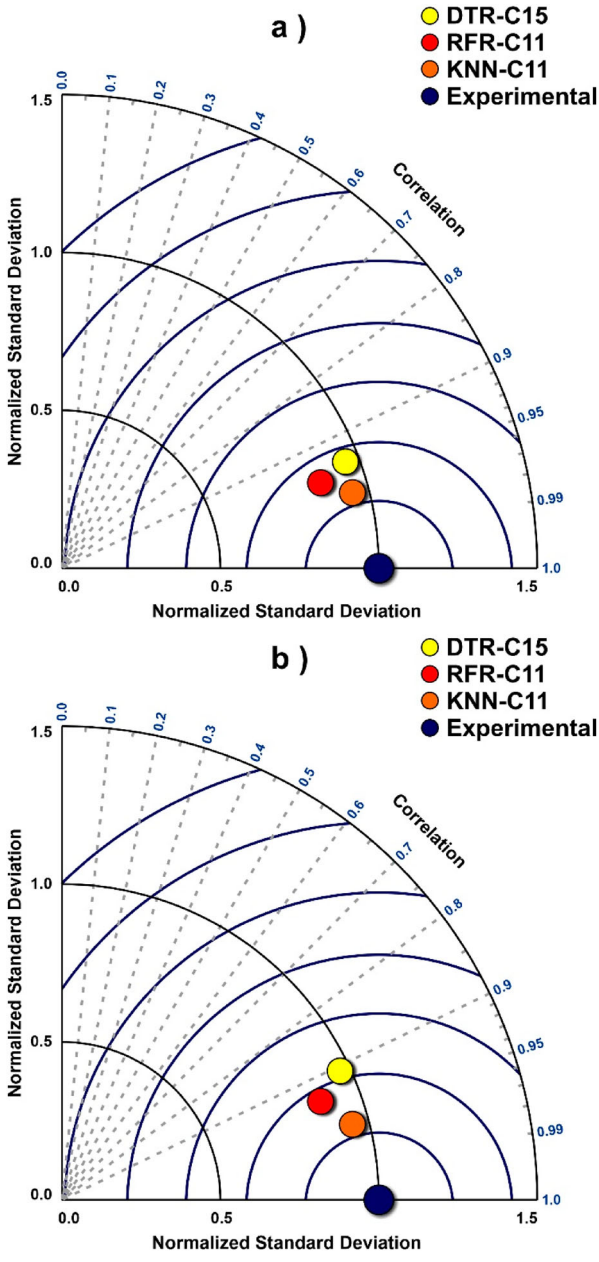

Figure 7. Taylor diagram of the predicted and experimental E_{20} values in (a) training dataset (b) testing dataset.

Table 9. Comparison of Performance indices between current study outcome and previous experimental results in testing stage.

Scholars	RMSE	R^2
Current study	0.003	0.929
(Preul & Holler, 1969)	0.959	0.460
(Avery & Novak, 1978)	0.041	0.322
(Markofsky & Kobus, 1978)	0.030	0.458
(Wormleaton & Tsang, 2000)	0.041	0.324

samples used in predictive models is divided into training and testing stages, with 163 and 74 samples, respectively. To assess the performance of developed models, few performance indices (i.e. RMSE, MAE, R^2 & WI) and some graphical representations (i.e. Agreement plot, Heat-map, Violin diagram, and Relative error plot) are used for both training and testing stages. Statistical results and graphical comparisons suggest that the KNN model with M11 combination ($MAE_{testing} = 0.002, R^2_{testing} = 0.929$), which includes W/L, S/L, and Fr input parameters is outperforming other models. The RFR-M11 ($MAE_{testing} = 0.003, R^2_{testing} = 0.886$) and DTR-M15 ($MAE_{testing} = 0.003, R^2_{testing} = 0.868$) model achieves the next rank for predicting (E_{20}) values at Parshall flumes, respectively. The input sensitivity analysis also indicates the Re and W/L as the least and most influential parameter on the prediction results, respectively. With relatively high-performance indices for all three predictive models and superiority to previous experimental results, this study reveals that the machine learning algorithms are highly accurate and versatile to solve the problems of hydraulic system engineering and management.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Ahmad Sharafati http://orcid.org/0000-0003-0448-2871

References

Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2020). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering*, *9*(1), 104599. https://doi.org/10.1016/j.jece.2020.104599

Avery, S. T., & Novak, P. (1978). Oxygen transfer at hydraulic structures. *Journal of the Hydraulics Division*, *104*(11), 1521–1540. https://doi.org/10.1061/JYCEAJ.0005100

Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chanson, H. (1995). Hydraulic design of stepped cascades, channels, weirs and spillways.

Chanson, H. (2002). *Hydraulics of stepped chutes and spillways*. CRC Press.

Cone, V. M. (1917). *The venturi flume*. US Government Printing Office.

Dursun, O. F. (2016). An experimental investigation of the aeration performance of parshall flume and venturi flumes. *KSCE Journal of Civil Engineering*, *20*(2), 943–950. https://doi.org/10.1007/s12205-015-0645-0

Emiroglu, M. E., & Baylar, A. (2003). An investigation of effect of stepped chutes with end sill on aeration performance. *Water Quality Research Journal*, *38*(3), 527–539. https://doi.org/10.2166/wqrj.2003.034

Ervine, D. A. (1998). Air entrainment in hydraulic structures: A review. *Proceedings of the Institution of Civil Engineers-Water Maritime and Energy*, *130*(3), 142–153. https://doi.org/10.14264/uql.2014.542

Gharib, R., Heydari, M., Kardar, S., & Shabanlou, S. (2020). Simulation of discharge coefficient of side weirs placed on convergent canals using modern self-adaptive extreme learning machine. *Applied Water Science*, *10*(1), 50. https://doi.org/10.1007/s13201-019-1136-0

Goel, E., Abhilasha, E., Goel, E., & Abhilasha, E. (2017). Random forest: A review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(1), 251–257. https://doi.org/10.23956/ijarcsse/V7I1/01113

Gulliver, J. S., Wilhelms, S. C., & Parkhill, K. L. (1998). Predictive capabilities in oxygen transfer at hydraulic structures. *Journal of Hydraulic Engineering*, *124*(7), 664–671. https://doi.org/10.1061/(ASCE)0733-9429(1998)124:7(664)

Heiner, B., & Barfuss, S. L. (2011). Parshall flume discharge corrections: Wall staff gauge and centerline measurements. *Journal of Irrigation and Drainage Engineering*, *137*(12), 779–792. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000355

Ho, T. K. (1995). Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. IEEE, pp 278–282.

Jović, S., Danesh, A. S., Younesi, E., Aničić, O., Petković, D., & Shamshirband, S. (2016). Forecasting of underactuated robotic finger contact forces by support vector regression methodology. *International Journal of Pattern Recognition and Artificial Intelligence*, *30*(7), 1659019. https://doi.org/10.1142/S0218001416590199

Kamiński, B., Jakubczyk, M., & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research*, *26*(1), 135–159. https://doi.org/10.1007/s10100-017-0479-6

Kaya, N., & Emiroglu, M. E. (2010). Study of oxygen transfer efficiency at baffled chutes. In: Proceedings of the Institution of Civil Engineers-Water Management. Thomas Telford Ltd, pp 447–456.

Kumar, M., Ranjan, S., Tiwari, N. K., & Gupta, R. (2018). Plunging hollow jet aerators-oxygen transfer and modelling. *ISH Journal of Hydraulic Engineering*, *24*(1), 61–67. https://doi.org/10.1080/09715010.2017.1348264

Kumar, M., Tiwari, N. K., & Ranjan, S. (2020). Prediction of oxygen mass transfer of plunging hollow jets using regression models. *ISH Journal of Hydraulic Engineering*, *26*(3), 291–300. https://doi.org/10.1080/09715010.2018.1485119

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R News*, *2*, 18–22. https://doi.org/10.1177/154405910408300516.

Markofsky, M., & Kobus, H. (1978). Unified presentation of weir-aeration data.

Mitchell, T. M. (1997). Machine learning.

Mohammed, M., Sharafati, A., Al-Ansari, N., & Yaseen, Z. M. (2020). Shallow foundation settlement quantification: Application of hybridized adaptive neuro-fuzzy inference system model. *Advances in Civil Engineering*, *2020*. https://doi.org/10.1155/2020/7381617

Mustonen, S. (1986). *Sovellettu hydrologia*. Helsinki.

Nikolić, V., Petković, D., Lazov, L., & Milovančević, M. (2016). Selection of the most influential factors on the water-jet assisted underwater laser process by adaptive neuro-fuzzy technique. *Infrared Physics & Technology*, *77*, 45–50. https://doi.org/10.1016/j.infrared.2016.05.021

Parshall, R. L. (1928). *The improved venturi flume*. CER.

Parshall, R. L. (1950). Measuring water in irrigation channels with Parshall flumes and small weirs. Circ (United States Dep Agric no 843.

Parshall, R. L., & Rohwer, C. (1921). The venturi flume. *Agricultural Experiment Station of the Agricultural College of Colorado*.

Petković, D., Ab Hamid, S. H., Ćojbašić, Ž, & Pavlović, N. T. (2014a). Adapting project management method and ANFIS strategy for variables selection and analyzing wind turbine wake effect. *Natural Hazards*, *74*(2), 463–475. https://doi.org/10.1007/s11069-014-1189-1

Petković, D., Gocic, M., Trajkovic, S., Milovančević, M., & Šević, D. (2017). Precipitation concentration index management by adaptive neuro-fuzzy methodology. *Climatic Change*, *141*(4), 655–669. https://doi.org/10.1007/s10584-017-1907-2

Petković, D., Shamshirband, S., Saboohi H., Ang, T.F., Anuar, N.B., Rahman, Z.A., & Pavlović, N.T. (2014b). RETRACTED: Evaluation of modulation transfer function of optical lens system by support vector regression methodologies–A comparative study.

Preul, H. C., & Holler, A. G. (1969). Reaeration through low dams in the Ohio River. In: Proceedings of the 24th Industrial Waste Conference-Part Two, Purdue University, Lafayette, Indiana.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(3), 221–234. https://doi.org/10.1016/S0020-7373(87)80053-6

Ranjan, S., & Tiwari, N. K. (2019). Aeration Efficiency Evaluation of Modified Small Parshall Flume Using M5P and Adaptive Neuro-Fuzzy Inference System. In: Sustainable Engineering. Springer, pp 243–252.

Reclamation USB of, Interior USD of the, Service USNRC. (1997). Water measurement manual. The Bureau.

Sangeeta, S. R. (2018). Aeration efficiency of small Parshall flume. National institute of technology Kurukshetra. India.

Sangeeta, S. R., & Tiwari, N. K. (2019). Aeration efficiency evaluation of modified small parshall flume using m5p and adaptive neuro-fuzzy. *Sustain Eng Proc EGRWSE*, *30*, 243–252. https://doi.org/10.1007/978-981-13-6717-5_24

Shabani, S., Samadianfard, S., Sattari, M. T., Mosavi, A., Shamshirband, S., Kmet, T., & Várkonyi-Kóczy, A. R. (2020). Modeling pan evaporation using Gaussian process regression K-nearest neighbors random forest and support vector machines; comparative analysis. *Atmosphere*, *11*(1), 66. https://doi.org/10.3390/atmos11010066

Sharafati, A., Haji Seyed Asadollah, S. B., Motta, D., & Yaseen, Z. M. (2020a). Application of newly developed ensemble machine learning models for daily suspended sediment load prediction and related uncertainty analysis. *Hydrological Sciences Journal*, *65*(12), 2022–2042. https://doi.org/10.1080/02626667.2020.1786571

Sharafati, A., Masoud, H., Tiwari, N. K., Bhagat, S.K., Al-Ansari, N., Chau, K.W., & Yaseen, Z.M. (2021). Performance evaluation of sediment ejector efficiency using hybrid neuro-fuzzy models. *Engineering Applications of Computational Fluid Mechanics*, *15*(1), 627–643. https://doi.org/10.1080/19942060.2021.1893224

Sharafati, A., Tafarojnoruz, A., Motta, D., & Yaseen, Z. M. (2020b). Application of nature-inspired optimization algorithms to ANFIS model to predict wave-induced scour depth around pipelines. *Journal of Hydroinformatics*, *22*(6), 1425–1451. https://doi.org/10.2166/hydro.2020.184

Sihag, P., Tiwari, N. K., & Ranjan, S. (2017). Modelling of infiltration of sandy soil using gaussian process regression. *Modeling Earth Systems and Environment*, *3*(3), 1091–1100. https://doi.org/10.1007/s40808-017-0357-1

Sihag, P., Tiwari, N. K., & Ranjan, S. (2019). Prediction of unsaturated hydraulic conductivity using adaptive neuro-fuzzy inference system (ANFIS). *ISH Journal of Hydraulic Engineering*, *25*(2), 132–142. https://doi.org/10.1080/09715010.2017.1381861

Sihag, P., & Vajesnayee, S. R. (2018). Prediction of trapping efficiency of vortex tube silt ejector. *ISH Journal of Hydraulic Engineering*. https://doi.org/10.1080/09715010.2018.1441752

Tiwari, N. K. (2019). Evaluating hydraulic jump oxygen aeration by experimental observations and data driven techniques. *ISH Journal of Hydraulic Engineering*, 1–15. https://doi.org/10.1080/09715010.2019.1658551

Tiwari, N. K., & Sihag, P. (2020). Prediction of oxygen transfer at modified Parshall flumes using regression models. *ISH Journal of Hydraulic Engineering*, *26*(2), 209–220. https://doi.org/10.1080/09715010.2018.1473058

Tiwari, N. K., Sihag, P., Singh, B. K., Ranjan, S. & Singh, K.K. (2019). Estimation of tunnel desilter sediment removal efficiency by ANFIS. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, *44*, 959–974. https://doi.org/10.1007/s40996-019-00261-3

Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761–1768. https://doi.org/10.1016/j.energy.2006.11.010

Wilhelms, S. C., Gulliver, J. S., & Parkhill, K. (1993). Reaeration at low-head hydraulic structures. *Army Engineer Waterways Experiment Station Vicksburg Ms Hydraulics Lab*.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record*, *31*(1), 76–77. https://doi.org/10.1145/507338.507355

Wormleaton, P. R., & Tsang, C. C. (2000). Aeration performance of rectangular planform labyrinth weirs. *Journal of Environmental Engineering*, *126*(5), 456–465. https://doi.org/10.1061/(ASCE)0733-9372(2000)126:5(456)

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Yarmohammadi, E., Yosefvand, F., Rajabi, A., & Shabanlou, S. (2019). Modeling discharge coefficient of triangular plan form weirs using extreme learning machine. *Journal of Applied Research in Water and Wastewater*, *6*(2), 80–87. https://doi.org/10.22126/arww.2019.1356