SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Comparison of Machine Learning Models in Forecasting Reservoir Water Level

Mohammad Amimul Ihsan Aquil[1], Wan Hussain Wan Ishak[2,3,*]

[1] Department of Computer and Information Sciences, Universiti Teknologi Petronas, Malaysia
[2] School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia
[3] Data Science Research Lab, School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

ARTICLE INFO

Article history:
Received 24 April 2023
Received in revised form 16 July 2023
Accepted 25 July 2023
Available online 10 August 2023

Keywords:
Deep learning; early water release decision; forecasting model; machine learning; reservoir water level

ABSTRACT

Reservoirs are important for flood mitigation and water supply storage. The reservoir water release decision, however, must be intelligently modeled due to the unknown volume of input. The model can help reservoir operators make early water release decisions during heavy rainstorms and hold water during drought seasons. One of the promising techniques has been a machine learning-based forecasting model. Therefore, in this study, several machine learning models were identified and compared in terms of performance using Mean Absolute Error (MAE), R-Square, and Root Mean Square (RMSE). The findings show that VARMAX has the highest R-squared value. This identifies the data set as a time series having a seasonal component. ARIMA, on the other hand, is unable to produce adequate results when a seasonal component is included. Both models' MAE and RMSE values accurately reflect the above-mentioned argument.

1. Introduction

Reservoir is a manmade or natural lake used to store water. Typically, reservoirs are built to provide a range of purposes, such as flood control, hydropower generating, water supply, recreation, and so forth. Besides being given high attention in global energy security, reservoirs are essential for lowering the danger of flooding in downstream communities during times of heavy rain [1]. During the dry season, however, the reservoir is used to keep the water supply for home usage. As a result, the reservoir operator must play a critical role in both seasons, making accurate decisions [2].

One of the decisions the operator must make is whether to discharge the water early. Early reservoir water release is necessary to ensure reservoir water storage for the water usages while also leaving enough space for incoming inflow [3,4]. The operator must rely on observing the real time reservoir water level and estimating the incoming inflow in order to make an informed judgement [5]. The incoming inflow is influenced by a variety of factors including the number of rivers upstream and the amount of rainfall recorded at different gauging stations. This is a challenging task that can

* *Corresponding author.*
E-mail address: wanhussain@gmail.com

https://doi.org/10.37934/araset.31.3.137144

only be performed by an experienced operator [6]. The decision to release reservoir water early can be made by the operator using the forecasted reservoir water level [7].

The forecasting model can be developed using machine learning algorithms. A machine learning algorithm is a computer programme that can be trained to recognize specific patterns and then apply what they've learned to make informed decisions [8]. A subset of machine learning called deep learning is more advanced model that can learn and make intelligent judgments on their own [9].

A deep learning model is developed to analyses data in real time using reasoning techniques like to those used by humans to reach decisions. Applications using deep learning do this by using an artificial neural network, which is a layered structure of algorithms [10]. Compared to other machine learning models, artificial neural networks have a much more powerful learning process since they are designed after the biological neural network of the human brain.

In this paper several machine learning algorithms were employed to learn a pattern of upstream rainfall at Timah Tasoh reservoir. Timah Tasoh reservoir is located in Perlis state which is one of the northern parts of Malaysia. The performance of each algorithm was compared based on the Mean Absolute Error (MAE), R-Square and Root Mean Square Error (RMSE).

2. Literature Review

Flood is one of the natural disasters that can cause huge economic losses. The patterns of flood occurrence can be utilized to advise the authorities on the best course of action and alert the community [11]. Hence, river and lake water level forecasting are essential. Phan and Nguyen [12] for instance employ time series hydrological prediction models to anticipate future events because water level data from hydrological stations has a structure that is comparable to a time series model. It is possible to unearth hidden information by forecasting reservoir water levels using historical data.

In modern dam impact studies, a major problem is the lack of a reliable model for projecting how changes in water level would affect reservoir operation [13]. Because some water must be released down the spillway to keep the water level below the full supply level, it reduces the efficiency of the reservoir [7,14].

Water level forecasting models may be developed using either a process-based numerical prediction technique or a data-driven approach. Process-based models typically deliver accurate outcomes and accurately capture the nature of physical processes [12]. However, because they need a variety of hydro-geomorphological parameters, including topography, geology, conduction roughness, and cross-section, they are computationally costly.

The data-driven methods consist of machine learning (ML), deep learning, and time series analysis. Yahya *et al.,* [15] study on the machine learning technique and application examines many ways an algorithm might model a problem in accordance with the available data or information. Selecting input parameters that will affect output parameters is one of the most crucial tasks in ML, since it requires focus and a full grasp of the underlying physical process based on causal factors and statistical analysis of prospective inputs and outputs.

Sapitang *et al.,* [14] explored different machine learning techniques such as decision forest regression (DFR), neural network regression (NNR), boosted decision tree regression (BDTR), and Bayesian linear regression (BLR) to identify the most accurate water level prediction model based on historical data that were measured daily from 1985 to 2019. From day one through day seven, several time frames were examined. The study demonstrated that every integrated model had positive outcomes and had the potential to mimic real values. But occasionally, anticipated values were shown to be close to the coefficient of determination of 1.

Phan and Nguyen [12] combined statistical machine learning model with AutoRegressive Integrated Moving Average (ARIMA) in water level forecasting. Their findings indicated that well-known and successful forecasting models that have been proposed and tested on hydrological time series include ARIMA, KNN, RF, SVM, and LSTM. The other statistical machine learning models are superior at modelling nonlinear time series, while ARIMA excels at modelling linear time series. A time series often has both linear and nonlinear correlation patterns in reality.

When analyzing the relationship between commercial factors, technique selection is challenging due to the special properties of time series data. The characteristics of time series that are most frequently seen include stationarity, trends, cycles, seasonality, autoregressiveness, and structural breakdowns. To strengthen the model, these characteristics must be addressed or accommodated [12].

Khai *et al.,* [16] proposed forecasting the use of the Time Series Regression model in conjunction with the Support Vector Machine model for predicting water levels. The proposed model can be used for reliable data predictions because of the lower error rate, and the correlation coefficient is closer to one. Reservoirs with competing set conditions, such as weather and hydrological conditions, can use the model to forecast hydrological states and increase model relevance.

In this paper, a fundamental paradigm for predicting water levels is presented, which will offer precise information for disaster prevention. In order to remove any mistakes that might result from utilizing only one of these methods, we examine the use of Machine Learning, Deep Learning, and Time Series Analysis together in this study.

2. Methodology

The experiment relied on datasets obtained from the Perlis Department of Irrigation and Drainage, Malaysia. The dataset's description is given in Table 1. The target variable is continuous in nature. Data analysis is carried out using Python programming, Scikit-learn (machine learning framework), Keras (deep learning framework), and Statsmodel (statistics model framework). Overall, there were 2,923 samples available in the dataset and they were split (70-30) for training and testing. Data was in good condition with no missing values. Date Code was removed during feature extraction and all categorical variables were labelled with multi-hot encoding.

Table 1
Description of the dataset

Variables	Description	Type
Date Code	Date of each day	Date
Day	Day number of each day	Integer
Month	Month number of each day	Integer
Year	Year number of each day	Integer
RF Total	Total Rainfall in each day	Decimal
RF Category	Category of rainfall in each day	Categorical
WL category	Category of water level e.g., Normal, Alert etc.	Categorical
Water Level	Level of water in reservoir each data	Decimal (Target Variable)

Twelve algorithms were chosen and employed in this study; (1) Linear Regression, (2) Passive Aggressive, (3) Decision Tree, (4) Random Forest, (5) Extra Tree, (6) Adaboost, (7) GradientBoost, (8) MVR, (9) LSTM Encoder-Decoder Model, (10) BI-LSTM, (11) ARIMA, and (12) VARMAX. The following measures were used to evaluate all of the algorithms outlined above:

i. Mean Absolute Error (MAE) – is a measure of errors between predicted and observed output.
ii. R-Squared - measures the strength of the relationship between the linear model and the dependent variables.
iii. Root Mean Square (RMSE) - is a standard way to measure the error of a model in predicting quantitative data.

3. Results and Discussion

The findings of this study are shown in Table 2. In time series applications, the LSTM model is useful. However, because BI-LSTM is a neural network, it requires a huge data set to train and perform well. As a result, the model performs poorly in our study because we only have a few data points to train it with.

Table 2
Results produce by the algorithms

Algorithms	MAE	R-Square	RMSE
1. Linear Regression	0.31	0.315	0.425
2. Passive Aggressive	0.367	0.012	0.51
3. Decision Tree	0.317	0.273	0.438
4. Random Forest	0.307	0.351	0.413
5. Extra Tree	0.316	0.286	0.434
6. Adaboost	0.312	0.391	0.4
7. GradientBoost	0.295	0.4333	0.386
8. MVR	0.303	0.417	0.392
9. LSTM Encoder-Decoder Model	0.213	0.95	0.329
10. BI-LSTM	0.452	0.41	0.573
11. ARIMA	0.426	-0.012	0.51
12. VARMAX	0.035	0.972	0.088

Gradient Boost classification method is 2nd best amongst all the algorithms in terms of R-squared values. Consequently, gradient boosting Regression creates a flimsy model that associates attributes with the residual. This procedure moves the model closer to the desired outcome by adding the residual predicted by a weak model to the input of the current model. Repeating this process increases the accuracy of the model. This is why it performs better than all other algorithms except VARMAX.

BI-LSTM employs inputs in two ways: forward and backward, which can help us improve performance in specific circumstances. In this study, the BI-LSTM outperforms the LSTM and many other traditional algorithms in terms of MAE and R-squared.

The findings also reveal that Adaboost is less flexible than Gradient boost, which is due to the fact that Gradient boost is a generic algorithm that aids us in obtaining a more approximate solution to additive modelling issues. In this study, the Adaboost performs poorly in comparison to the Gradient boost, but it outperforms other classical methods since it effectively minimises the loss function associated with any classification error.

Random forest, unlike Gradient boost, does not begin the process of integrating the decision trees at the beginning, but rather at the end. Although random forest constructs each tree individually, Gradient boost constructs each tree one at a time. As a result, Gradient boost outperforms Random forest in the vast majority of scenarios. Random forest, on the other hand, outperforms linear regression in this scenario because random forest better captures non-linearity than linear regression.

This study's findings also demonstrate that decision trees aren't working well. This is most likely owing to concerns like overfitting, bias, and variation. Because they function on a simple decision-making process, they don't always perform as well as the advanced versions of Decision tree, such as Random Forest and Gradient boost.

The Extremely Randomized Trees Classifier (Extra tree) is quite similar to the random forest tree, however, it creates decision trees differently. The relevance of each feature is determined here, and the user then chooses the amount of features depending on his preferences. In this study, Extra tree produce R-squared 0.286 and MAE 0.316. Considering the fact that the dataset is very small, the user has very limited options in choosing the features.

In contrast to batch learning, when the complete dataset is processed at once, the Passive Aggressive method is an online-learning algorithm where the input data comes in sequential order. Every time a new input is received, the model is updated. This algorithm is generally used where the huge amount of data needs to be analyzed. However, because the data set is limited in this study, it does not allow it to learn enough from the data and hence performs poorly in terms of R-squared and errors.

ARIMA (Auto Regressive Integrated Moving Average) is a technique for forecasting future values based on previous values in a time series. The findings shows that ARIMA performance in terms of R-squared and Errors are poor. That's because the dataset has some elements that ARIMA's auto regressive and moving average components can't effectively fit or explain in a model.

The results also demonstrate that MVR is not a good fit for this dataset. In comparison to other competitive high-performance algorithms, the MVR performs poorly. Despite this, it outperforms other machine learning-based techniques.

The plot of MAE for all of the methods utilised in this study is shown in Figure 1. In the case of MAE, a lower value is preferable. MAE is a statistic that assesses the average magnitude of mistakes in a set of forecasts without taking into account their direction (positive or negative). It is the average of the absolute differences between prediction and actual observation over the test sample, where all individual deviations are given equal weight. The lowest MAE value is found in VARMAX, while the highest value is found in BI-LSTM. While many of the algorithms have MAE values in the range of 0.3, many of them are extremely comparable.

The plot of R-Squared for all of the algorithms can be seen in Figure 2. When compared to those with a lower value, those with a higher value suggest better performance. R-Squared has a constant range of values between 0 and 1. In a regression model, R-Squared is a statistical measure of fit that shows how much variance in a dependent variable is explained by the independent variable(s). In this situation, R-squared can be defined as the percentage of a 4 dependent variable's (RF category) movements that can be explained by independent variable movements (Water level, WL category, RF total). The VARMAX model has the highest R-squared value, as shown in Figure 2.

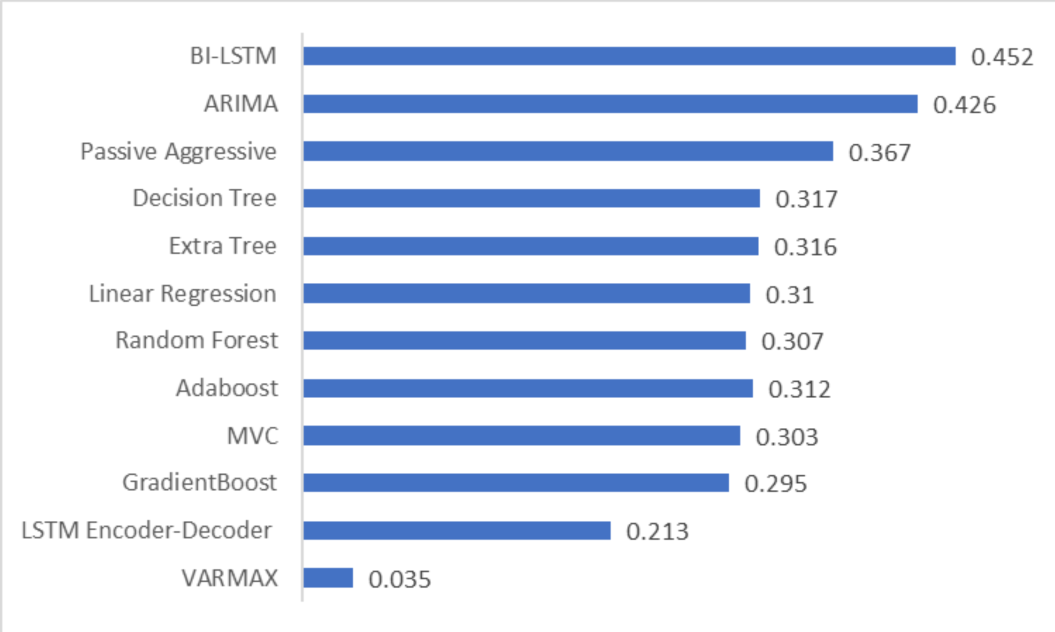

Fig. 1. Comparison of MAE value for each algorithm

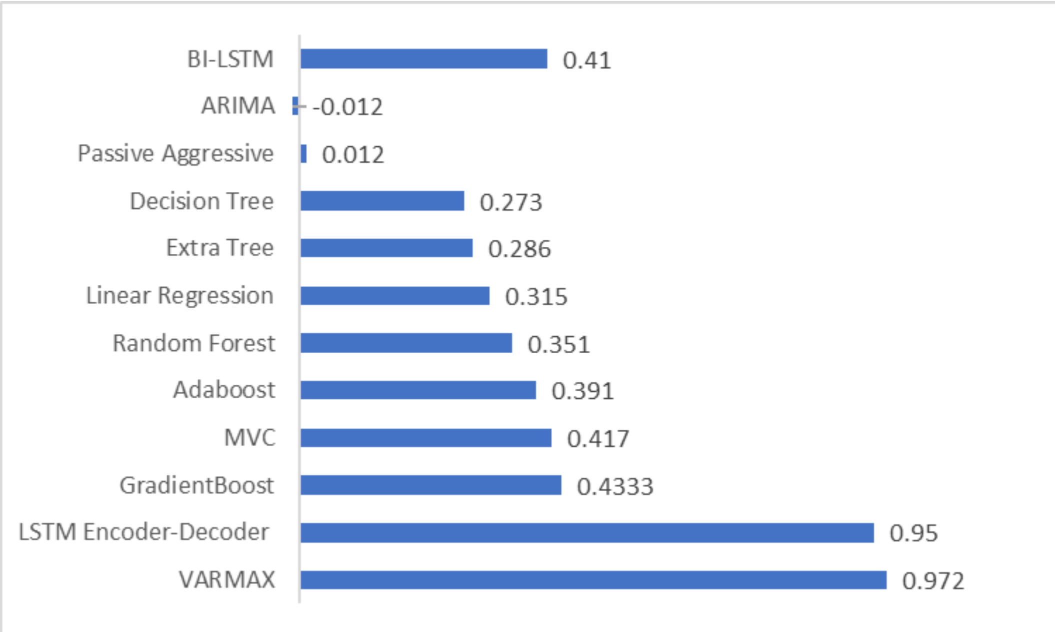

Fig. 2. Comparison of R-Square value for each algorithm

The graph in Figure 3 shows the RMSE plot for all algorithms. In this case, a lower RMSE value indicates higher performance. The RMSE is a quadratic scoring rule that additionally calculates the average magnitude of the error. It's the square root of the average of squared differences between predicted and observed results. In this study, VARMAX has the lowest RMSE value, indicating that it has the best performance of all the algorithms used. Other algorithms their RMSE is in the range of 0.3 to 0.58. The RMSE should be more useful as compared to the MAE, when large errors are particularly undesirable.

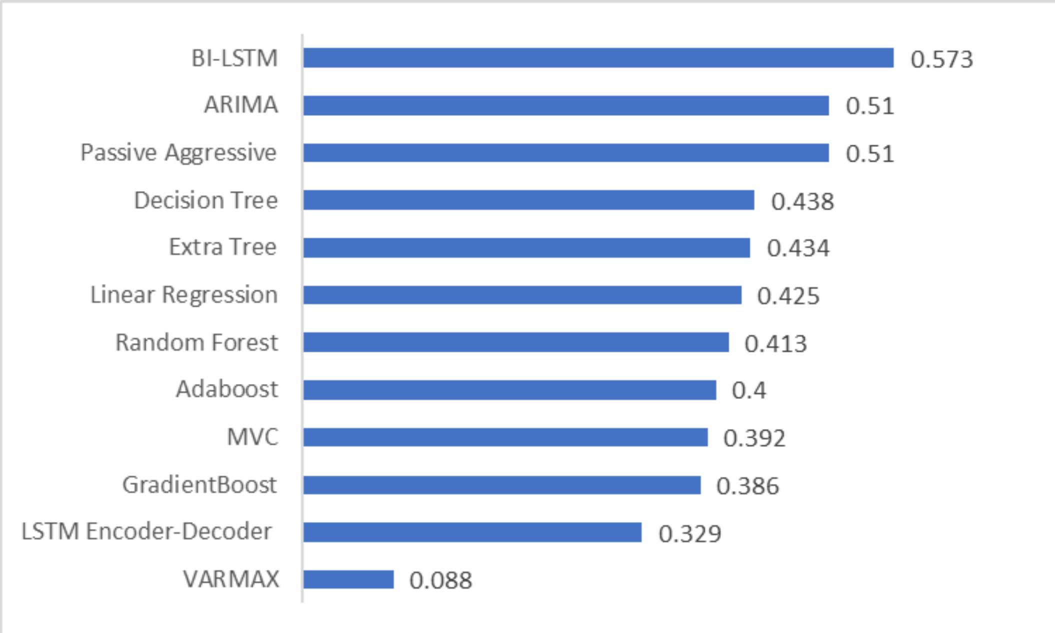

Fig. 3. Comparison of RMSE value for each algorithm

4. Conclusions

The results of this study reveal that machine learning models perform differently on time series data having a seasonal component. The findings show that VARMAX was found to be the best machine learning model, having the highest R-squared value and the lowest MAE and RSME values. ARIMA, on the other hand, is unable to give adequate results. VARMAX has the potential to be integrated into a reservoir water level forecasting system, which could help reservoir operators make early water release decisions. This approach will benefit the reservoir operation as the conventional approach in decision making was constrained due to the large amount of data and operator's experience [5].

Acknowledgement
This research was not funded by any grant.

References
[1] Khattak, Muhammad Adil, Lee Jun Keat, Khairul Anwar Bapujee, Tan Xin Hui, Amirul Syafiq Othman, Afiq Danial Abd Rasid, Lailatul Fitriyah Ahmad Shafii, and Suhail Kazi. "Global energy security and Malaysian perspective: A review." *Progress in Energy and Environment* 6 (2018): 1-18.

[2] Giudici, Federico, Daniela Anghileri, Andrea Castelletti, and Paolo Burlando. "Descriptive or normative: How does reservoir operations modeling influence hydrological simulations under climate change?." *Journal of Hydrology* 595 (2021): 125996. https://doi.org/10.1016/j.jhydrol.2021.125996

[3] Hassin, Mohd Hafiz B. Mohd, Norita Bt Md Norwawi, and Azizi B. Ab Aziz. "Temporal Case-Based Reasoning for reservoir spillway gate operation recommendation." In *2006 International Conference on Computing & Informatics*, pp. 1-4. IEEE, 2006. https://doi.org/10.1109/ICOCI.2006.5276517

[4] Ma, Chao, Rui Xu, Wei He, and Jinjin Xia. "Determining the limiting water level of early flood season by combining multiobjective optimization scheduling and copula joint distribution function: A case study of three gorges reservoir." *Science of The Total Environment* 737 (2020): 139789. https://doi.org/10.1016/j.scitotenv.2020.139789

[5] Wan Ishak, Wan Hussain, Ku Ruhana Ku-Mahamud, and Norita Md Norwawi. "Modelling of human expert decision making in reservoir operation." *Jurnal Teknologi* 77, no. 22 (2015): 1-5. https://doi.org/10.11113/jt.v77.6654

[6] Zhu, Feilin, Ping-an Zhong, Bin Xu, Juan Chen, Yimeng Sun, Weifeng Liu, and Tiancheng Li. "Stochastic multi-criteria decision making based on stepwise weight information for real-time reservoir operation." *Journal of Cleaner Production* 257 (2020): 120554. https://doi.org/10.1016/j.jclepro.2020.120554

[7] Wan Ishak, Wan Hussain, Ku Ruhana Ku-Mahamud, and Norita Md Norwawi. "Intelligent decision support model based on neural network to support reservoir water release decision." In *Software Engineering and Computer Systems: Second International Conference, ICSECS 2011*, Kuantan, Pahang, Malaysia, June 27-29, 2011, Proceedings, Part I 2, pp. 365-379. Springer Berlin Heidelberg, 2011. https://doi.org/10.1007/978-3-642-22170-5_32
[8] Dey, Ayon. "Machine Learning Algorithms: A Review." *International Journal of Computer Science and Information Technologies* 7, no. 3 (2016): 1174-1179.
[9] Alaskar, Hind, and Tanzila Saba. "Machine Learning and Deep Learning: A Comparative Review." *Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020* (2021): 143-150. https://doi.org/10.1007/978-981-33-6307-6_15
[10] Shrestha, Ajay, and Ausif Mahmood. "Review of deep learning algorithms and architectures." *IEEE Access* 7 (2019): 53040-53065. https://doi.org/10.1109/ACCESS.2019.2912200
[11] Ku-Mahamud, Ku Ruhana, Norharyani Zakaria, Norliza Katuk, and Mohamad Shbier. "Flood pattern detection using sliding window technique." In *2009 Third Asia International Conference on Modelling & Simulation*, pp. 45-50. IEEE, 2009. https://doi.org/10.1109/AMS.2009.15
[12] Phan, Thi-Thu-Hong, and Xuan Hoai Nguyen. "Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river." *Advances in Water Resources* 142 (2020): 103656. https://doi.org/10.1016/j.advwatres.2020.103656
[13] Ashaary, Nur Athirah, Wan Hussain Wan Ishak, and Ku Ruhana Ku-Mahamud. "Neural network application in the change of reservoir water level stage forecasting." *Indian Journal of Science and Technology* 8, no. 13 (2015): 1-6. https://doi.org/10.17485/ijst/2015/v8i13/70634
[14] Sapitang, Michelle, Wanie M. Ridwan, Khairul Faizal Kushiar, Ali Najah Ahmed, and Ahmed El-Shafie. "Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy." *Sustainability* 12, no. 15 (2020): 6121. https://doi.org/10.3390/su12156121
[15] Yahya, F., B. Fazli, H. Sallehudin, and M. Jaya. "Machine learning in dam water research: An overview of applications and approaches." *International Journal of Advanced Trends in Computer Science and Engineering* 9, no. 2 (2020): 1268-1274. https://doi.org/10.30534/ijatcse/2020/56922020
[16] Khai, Wong Jee, Moath Alraih, Ali Najah Ahmed, Chow Ming Fai, Ahmed El-Shafie, and Amr El-Shafie. "Daily forecasting of dam water levels using machine learning." *International Journal of Civil Engineering and Technology (IJCIET)* 10, no. 6 (2019): 314-323.