A Performance Comparison of Various Artificial Intelligence Approaches for Estimation of Sediment of River Systems

Gasim Hayder1,2, Mahmud Iwan Solihin3, Khairul Faizal Bin Kushiar4

1 Institute of Energy Infrastructure, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, Malaysia
2 Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, Malaysia
3 Faculty of Engineering, Technology and Built Environment, UCSI University, Jalan Puncak Menara Gading, Taman Connaught, Kuala Lumpur 56000, Malaysia
4 Asset Management Department, Generation Division, Tenaga Nasional Berhad, 59200 Kuala Lumpur, Malaysia
* Corresponding author’s e-mail: gasim@uniten.edu.my

ABSTRACT
Sediment is a universal issue that is generated in the river catchment and affects the river flow, reservoir capacity, hydropower generation and dam structure. This paper aims to present the result of experimentation in sediment load estimation using various machine learning algorithms as a powerful AI approach. The data was collected from eight locations in upstream area of Ringlet reservoir catchment. The input variables are discharge and suspended solid. It was found that there is strong correlation between sediment and suspended solid with correlation coefficient of $R = 0.9$. The developed ML model successfully estimated the sediment load with competitive results from ANN, Decision Tree, AdaBoost and SVM. The best result was produced by SVM (v-SVM version) where very low RMSE was generated for both training and testing dataset despite its more complicated hyper-parameters setup. The results also show a promising application of machine learning for future prediction in hydro-informatic systems.

Keywords: sediment estimation, artificial intelligence, machine learning prediction, river systems, visual programming, hydro-informatics systems

INTRODUCTION

The information of sediment load applied in designing reservoirs and dams, sediment transportation and pollutants in rivers and lakes, designing stable channels and dams, preservation of aquatic life and wildlife, watershed management, and assessment for environmental impact (Cigizoglu, 2004). The nature of suspended sediment is non-linear, which necessities application of a non-linear method for the forecasting and predicting study. The prediction of sediment is critical because it tends to affect the hydraulic river structures (Kisi et al., 2009; Kisi, 2010). Sediment transport is one of the most significant issues in surface water resources (Gericke and Venohr, 2012). In the previous study, rainfall and stream flow were reported as primary factors that influenced the suspended load (Jie-Lun and Yu-Shuie, 2011).

These phenomena attract researcher attention to develop direct and also indirect simulation and prediction models that could be accepted by operators worldwide; however, there is a demand to look into each catchment for better forecasting (Abrahart et al. 2008 and Khan et al. 2019). The estimation of suspended sediment is challenging because it is closely related to flow and the mechanism of their non-linear relationship and their complex interactions with each other. Several kinds of research were conducted to estimate and understand the mechanism sediment
concentrations and sediment movement in the natural rivers, using various computing methods (Demirchi et al. 2015 and Mustafa & Isa, 2014). Establishment of a rating relationship between flow and suspended sediment concentration is a non-linear mapping. The commonly used statistical tools include curve fitting and regression. With the complexity of the issue, these techniques are not enough. There are techniques available for time series analysis, assuming a linear relationship between the variables. However, temporal changes in data exhibit complex non-linear behaviour and impose difficulties for accurate prediction. Therefore, a non-linear model using artificial intelligence can capture the complex temporal variations in time series data, is user-friendly and can produce results faster than most conventional models, as mentioned by Khanchoul et al. (2015).

The objective of this paper was to develop sediment load estimation from available input predictor variables using various machine learning (ML) algorithms as a powerful AI approach. The ML algorithms have been applied in numerous hydrology studies such as in (Hayder et al., 2020) where ANN is used. Some studies also used ANN particularly for sediment prediction, such as (Olyaie et al., 2015) (Afan et al., 2014) (Nivesh & Kumar, 2017).

METHODOLOGY

Study Area

Cameron Highland, well known to all as one of the tourist hotspots that is rich in active highland agriculture such as vegetables and tea farms, is situated in Pahang State, peninsular Malaysia (Razad et al., 2018) as shown in Fig. 1 and Fig. 2. Situated in the north-west State of Pahang, it bordered by another two states which are Kelantan on the north side and Perak on the west side and also bordered by Lipis district on the South-East side of the Cameron Highlands district (Razali et al., 2020a, Pradhan & Lee, 2010). This unique district has a total area roughly in 69,699 km² with a height between 300 m and 2,016 m above sea level. Moreover, the Cameron Highlands district also situated in the main range of Banjaran Titiwangsa and the highest temperature recorded in this district was 12°C (How Jin Aik et al., 2020, Gasim et al., 2009a).

There are six water catchments within the area of Cameron Highlands, namely Mensun, Lemoi, Wi, Terla, Telom and Bertam. Furthermore, several towns are located in the district such as Kampung Raja, Kuala Terla, Kea Farm, Brinchang, Teringkap, Tanah Rata and the southernmost town in the district, Ringlet (Gasim et al., 2009b). Sultan Abu Bakar (SAB) Dam and Ulu Jelai Dam are the major impoundments in the Cameron Highlands district (Razali et al., 2020b). With the height of 39.6 m and length of 135 m, the concrete SAB Dam was built with an amount of 19,000 m³ of rockfill and 52,000 m³ of concrete. Another name for SAB Dam is Ringlet Dam and this dam was built in the 1960s (Kun & Saman, 2004, Luis et al., 2013).

In addition, due to the construction of SAB Dam, a lake that is known as Ringlet Reservoir was produced. As an intrinsic part of the Cameron Highlands Hydroelectric Scheme of the National Electric Board, Ringlet Reservoir produces twenty per cent of total power in West Malaysia (Tayebiyan et al., 2014). The reservoir is approximately 0.4 km wide and 3.2 km long and it is operated by Tenaga National Berhad (TNB), one of the national utility companies in Malaysia. Moreover, the reservoir impounds the combined of several rivers, namely Kial, Kodol, Plau’ur and Telom Rivers which have been redirected from Telom water catchments through a tunnel (Telom Tunnel) into the Bertam catchments and also the Bertam River and its tributaries (Tayebiyan et al., 2016, Faudzi et al., 2019). Apart from being a hydro-power generator for Jor Power Plant, Ringlet Reservoir also acts as a flood control to protect the people who live in the Bertam Valley from flood damage. Installed and managed by Tenaga Nasional Berhad (TNB), Metrological Department of Malaysia (MMD) and the Department of Irrigation and Drainage (DID), it has shown that the Cameron Highlands catchment is equipped with huge hydrological network including stream flow, rainfall and weather stations (Razad et al., 2020a, Razad et al., 2020b).

The locals are taking advantages of the cold climate at Cameron Highlands to make incomes from two main sectors, namely tourism and agriculture. However, landslide and soil erosion occur instead due to the advanced development which increases in land use (Maturidi et al., 2020). The activities may also have indirectly or directly reduce the river water quality, thus
affecting the water quality of the Ringlet Reservoir (Jamil et al., 2014). Furthermore, the storage volume of water in the reservoir is always used to the fullest extent for generation of hydropower and to control flood. Nonetheless, due to sedimentation, the reservoir storage volume is gone and the energy output from the power station will be affected. The worst scenario would happen if the Ringlet Reservoir slowly lost its ability to hold huge flood inflows and therefore, inevitable control over the release of flood discharge through the spillway would occur (Luis et al., 2012).

Machine Learning Algorithms

Machine learning (ML) has attached much attention nowadays, due to emergence of success of deep learning applications, especially in computer vision. ML is subfield of AI approach that learns the behaviour of pattern from data to make prediction of output. In other words, machine learning a data-driven predictive modelling. Often, ML works in the same way as statistical learning, because it is also a data-driven approach. Classical statistical learning that can be categorized as basic ML is, for example, multiple linear regression (MLR) and also multiple non-linear regression (MLNR).

There are various ML algorithms available that some might not have attracted huge attention in applied research despite their performance. Some ML algorithms that were applied in this paper to estimate the sediment load in the river systems are: multiple linear regression (MLR), artificial neural networks (ANN), support vector machine (SVM), decision tree (DT), random forest (RF), and adaptive boosting/AdaBoost (AB). Some fundamental concepts of ML can be found in many resources, including those available online such as (Machine Learning 101 – Medium, n.d.). The mathematical detailed discussion of each algorithm is beyond the scope of this paper.

One thing that is common in ML predictive model building, is the algorithm training prior to the ML model validation/testing and deployment. Generally, the following flowchart in Figure 2 shows the model development workflow using the AI-machine learning approach which begins with data collection and preparation as well as ends with model deployment, monitoring and updating. This paper presents the middle stage of experimentation to train the various ML model using training dataset and validate the accuracy using testing dataset.
Data Collection and Preparation

The main ingredient of this research is the data itself. The data recorded was originally obtained from the dam operator. The raw data consists of records from input variables namely discharge in mg/L (DC), suspended solid in cumecs (SS) and output/target variable namely sediment in ton/day unit. From the initial observation of the raw data recorded, extensive data cleansing needs to be performed such as imputing missing data and exploratory data analysis. This task is normally a time-consuming and tedious effort in building a machine learning model. After the process, the available data with two input variables (DC and SS) in eight locations are compiled. The eight locations surrounding the river systems in Ringlet reservoir are listed in Table 1. The total number of data compiled is 415 data instances scattered on random days from 12 December 1997 to 12 May 2010 with only 1 sediment data missing. Prior to the training, the missing value was replaced with the average and the input features are normalized to the range of [0 to 1].

Table 1 summarizes the data used in the ML prediction model building. Furthermore, these 415 data instances are partitioned with random sampling method to split the data into a training dataset and a testing dataset with ratio of 80%-20% resulting 332 and 83 data instances for the training and the testing dataset, respectively.

Parameter Setup of The Regression Algorithms

In the ML model building, the task is basically to find the optimum configuration of the model involving the architecture and the hyper complicated parameters setup prior to the training, while some involve very simple setup or none whatsoever. In the applied research of this paper, the ML parameters setup and the benchmarking algorithms setup, i.e., MLR and MNLR, is summarized in Table 3 provided according to the respective algorithms. In this case, two MNLR (MNLR-1 and MNLR-2) are used to have different non-linear model equations. MLNR-1 model has been inspired by another study in sediment prediction (Olyaie et al., 2015). The implementation of the experimentation was carried out using free Orange software ver-3.25 with visual programming approach (Orange – Data Mining Fruitful & Fun, n.d.). This setup was obtained after some extensive experimentation with experienced initial guess and trial-and-error tuning.

Model Evaluation and Correlation Analysis

The task in this research is basically building a nonlinear multi-variable regression model except for MLR. For the benchmarking purpose,

Table 1. Locations with available data in the river systems (Ringlet reservoir)

<table>
<thead>
<tr>
<th>No.</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sungai Plaur at Outfall</td>
</tr>
<tr>
<td>2</td>
<td>Sungai Telom at Batu 49</td>
</tr>
<tr>
<td>3</td>
<td>Sungai Bertam at Robinsons Fall</td>
</tr>
<tr>
<td>4</td>
<td>Sungai Kial Above Intake</td>
</tr>
<tr>
<td>5</td>
<td>Sungai Ringlet at Indian Temple</td>
</tr>
<tr>
<td>6</td>
<td>Sungai Habu above Kuala Habu</td>
</tr>
<tr>
<td>7</td>
<td>Sungai Anak Ulong at The Boh Estate</td>
</tr>
<tr>
<td>8</td>
<td>Sungai Bertam above Ringlet Reservoir</td>
</tr>
</tbody>
</table>
the results of ML model were also compared with conventional MLR model and MNLR as used by (Olyaie et al., 2015) for sediment prediction. The developed model after the training process was evaluated in terms of the accuracy (as illustrated in Figure 2). The metrics of evaluation are regression coefficient (R²) and root mean squared error (RMSE) expressed as:

\[ R^2 = 1 - \frac{\text{MSE}}{\text{variance}(y_i)} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \]

\[ \text{NMSE} = \frac{\text{MSE}}{\text{variance}(y_i)} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \sum_{i=1}^{n}(y_i - \bar{y})^2 \]

\[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \]

where: \( \hat{y}_i \) denotes the predicted value,
\( y_i \) is the actual value
\( \bar{y}_i \) is the mean. NMSE is normalized mean squared error (MSE).

In addition, the correlation analysis was also performed to see how the target variable was correlated with the input variables. This analysis is often useful in selection of input variables/features especially when many number of input variables involved. The correlation coefficient (R) between two variables \( x \) and \( y \) can be calculated by dividing the covariance with the product of the standard deviations (\( \sigma \)) of the two variables as follows:

\[ R = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} \]

**RESULTS AND DISCUSSION**

The experimentation in Orange software follows the concept of visual programming that is compiled in a workflow. The workflow consists of blocks (called widgets) connection, as illustrated in Figure 3.

The widgets workflow shown in Figure 2 includes the correlation analysis result. On the basis of the data, the correlation analysis shows strong correlation between target variable (sediment) and the SS (R=0.9) and weak correlation to DC (R=0.24). This strong correlation between sediment and SS indicates the potential success of predictive model building for sediment using SS as predictor in addition to DC.
With the setup of the ML parameters listed in Table 3, the experimentation results can be compiled from the prediction widget. The results of the ML model performance is shown in Table 4. The results shows good performance for all ML model. The worse performance is undoubtedly produced by conventional MLR which is a linear regression method and MNLR despite its deterministic nonlinearity. The result of the rest non-linear ML algorithms produce $R^2$ higher than 0.94 both in training and testing. The best result is produced by SVM where very low RMSE is generated for training and testing dataset. Figure 4 shows the estimation result of SVM on test dataset where the error line (green) is flat at all instances of test dataset.

Lastly, the workflow of the AI-based predictive model development in this paper is not only useful for the local case study area. It can be useful as a modelling approach in other river systems. If the approach is applied in other area, the main task of the project to collect the data for predictive model building.

**CONCLUSIONS**

This research presents the experimentation results of sediment load prediction using various machine learning algorithms with two input variables, namely discharge and suspended solid. The data used in the case study was obtained in river systems surrounding Ringlet reservoir in Cameron highlands, Malaysia. The main finding indicates a promising result of sediment prediction using machine learning approaches. However, the developed ML model framework in this paper must be validated further using larger dataset in the next phase of data collection before being deployed. In this experimentation, SVM shows the outstanding performance as compared to the rest despite its more complicated parameter setup that must be done carefully during the model building. In addition, the performances of ANN, Decision Tree and AdaBoost are very competitive to that of SVM.

Furthermore, the antecedence values of input variables can also be used as predictor variables.

**Table 4. ML model performance**

<table>
<thead>
<tr>
<th>ML model</th>
<th>$R^2$ – train</th>
<th>RMSE – train</th>
<th>$R^2$ – test</th>
<th>RMSE – test</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR</td>
<td>0.88</td>
<td>10.65</td>
<td>0.88</td>
<td>13.12</td>
</tr>
<tr>
<td>MNLR-1</td>
<td>0.88</td>
<td>22.5</td>
<td>0.76</td>
<td>24.58</td>
</tr>
<tr>
<td>MNLR-2</td>
<td>0.89</td>
<td>22.0</td>
<td>0.72</td>
<td>26.24</td>
</tr>
<tr>
<td>ANN</td>
<td>0.99</td>
<td>5.18</td>
<td>0.97</td>
<td>6.69</td>
</tr>
<tr>
<td>DT</td>
<td>1</td>
<td>0</td>
<td>0.97</td>
<td>6.48</td>
</tr>
<tr>
<td>RF</td>
<td>0.94</td>
<td>2.33</td>
<td>0.96</td>
<td>7.43</td>
</tr>
<tr>
<td>AB</td>
<td>1</td>
<td>0.05</td>
<td>0.99</td>
<td>4.14</td>
</tr>
<tr>
<td>SVM*</td>
<td>0.99*</td>
<td>0.46</td>
<td>1*</td>
<td>0.19*</td>
</tr>
</tbody>
</table>
as these values can affect the target variable values. If dredging of sediment at downstream is carried out, i.e. Ringlet Weir in this case, this variable and its antecedent can also be used as predictor variables.

In the deployment stage, the selection of suitable ML model should consider some aspects such as human interpretability of the model and feasibility of the model deployment from the perspective of computational cost. Furthermore, deployment of the ANN-based model as an AI approach is of future interest in the scenario digitalization in hydro-informatic systems and involvement of IoT (Internet of things).

Acknowledgement

The authors would like to acknowledge the support from Tenaga Nasional Berhad (TNB), Universiti Tenaga Nasional (UNITEN), and UNITEN R&D Sdn. Bhd. (URND) under seed funding project.

REFERENCES

4. Bahagian Sumber Air dan Hidrology, JPS Malaysia


