ISSN 2278-3091

Volume 8, No.1.4, 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse2381.42019.pdf

https://doi.org/10.30534/ijatcse/2019/2381.42019

Comparing Performance of Machine Learning Algorithms in a Flood Prediction Model with Real Data Sets

Fadratul Hafinaz Hassan[1*], Nur Amiera Azelan[1]

[1]School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia, fadratul@usm.my*

ABSTRACT

Flood is one of an unforeseen and often sudden event or a situation that causes severe damage, destruction and human suffering which requires help by requesting to national or international level. The implementation of machine learning approaches in flood prediction may reduce all the risk factors. Machine learning is one of the method that provide better performances and it is cost-effective and recently used among hydrologists. However, the capability of each machine learning algorithm is different for each type of tasks which is called generalization problem. Thus, for this research, three machine learning methods which are Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) is chosen for flood prediction model. Each of machine learning algorithms are built and trained in order for they to work accordingly with two different datasets. The aim of this research project is to investigate the performance of three selected machine learning algorithms and compared their accuracy. ANN has shown promising results with the highest performance accuracy of 98% in dataset 1 and 77.10% in dataset 2.

Key words: Artificial Neural Network, Flood Prediction Model, Machine Learning Algorithms, Real Data Sets

1. INTRODUCTION

Flood is one of an unforeseen and often sudden event or a situation that causes severe damage, destruction and human suffering which requires help by requesting to national or international level. A flood disaster management which defined as a systematic process with the aim of reducing the negative impact or consequences to safe people and also infrastructure is applied when disaster hits. The main reason flood is chosen is because it is one of the natural disaster that often hits Southeast Asia. According to [1], floods in Southeast Asia are currently affecting 9.6 million people with 5.3 million in Thailand alone. This is because the region of Southeast Asian is prone to floods which are caused by heavy monsoon showers, typhoons and storms. Most recent floods have occurred in the mainland and island of Penang, Malaysia due to a 17-hour storm and in Vietnam due to Typhoon Damrey.

In this research project, flood prediction is focused more as it plays a major role in reducing risk, loss of human life and property damage in the future. Machine learning method is applied for flood prediction to work efficiently. Machine learning is one of the method that provide better performances and it is cost-effective. The application of machine learning method is quite famous as it is mostly used among hydrologists.

Machine learning is an advanced data-driven model which is a field of artificial intelligence (AI) is applied to get consistent results based on historical data. It works by extracting the information from the historical and past data and use it to predict the patterns of the trends and behavior. Its implementation is easier as it trains the model faster with low computation cost. Data-driven models is quicker to develop even with minimal inputs. The validation, testing and evaluation of data-driven models compared to physical models is also less complex and high in performance. Based on [2], the application of machine learning over the past 20 years demonstrated their suitability for flood prediction.

Thus, for this research project, three machine learning method which are Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) is chosen in flood prediction and discussed further in this research project. Machine learning is introduced as it analyzes data faster especially big data without having programmers to do it manually. Manually interpretation and analysis of integrated data is not relevant anymore. In this research project, the accuracy performance of Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) is studied and compared. Machine learning algorithm that produces the most accurate reading in flood prediction will be discussed further.

2. PREVIOUS WORKS OF MACHINE LEARNING MODELS

Machine learning techniques which requires machines to be trained to detect damages will help in reducing the interaction of human which will improve the performance of decision making. Most of machine learning models and algorithms improves the accuracy of performance compared to traditional classification and change detection methods.

To make the process of disaster management works in an effective and efficient way, machine learning was introduced for the past 20 years. Machine learning and big data analytics works together in improving in flood modelling and prediction [3]. Flood prediction plays a major role in reducing risk, loss of human life and property damage in the future. Machine learning method is applied for flood prediction to work efficiently. Machine learning provide better performances in prediction and it is cost-effective. The application of machine learning method is quite famous as it is mostly used among hydrologists.

According to [2], in the wake of considering strategy of machine learning methods in most of the flood-related events, Multilayer Perceptron (MLP), ANNs, DT, SVM, Adaptive Neuro-Fuzzy Inference System (ANFIS), Wavelet Neural Network (WNNs) and finally Ensemble Prediction Systems (EPSs) are the most well-known techniques. Multilayer perceptron (MLPs) are a portrayal of ANNs which as of late picked up ubiquity.

Each of the machine learning models involves different algorithms. Among all machine learning models, ANN is said to be the foremost known learning algorithms in modeling flood prediction. In comparison to traditional statistical models, the ANN approach was used for prediction with greater accuracy as it derives the meaning from past and historical data. SVM is greatly popular in flood modelling which today known as robust and efficient machine learning algorithm for flood prediction. DT are classified as fast algorithms, they became very popular in ensemble forms to model and predict floods.

2.1 Flood Prediction Models using Artificial Neural Network, Support Vector Machine and Decision Tree

[4], [5] implemented ANN to predict flood prediction by predicting maximum daily flow and daily pan evaporation. The collected data included rainfall, temperatures, humidity, and sunshine hours from the same watershed meteorological stations of subtropical climates. However, both works only tested on one data set or only used data set from the study case area. [6] compared ANN and SVM by proposing two time-series models in predicting the variances of the level of ground water. The type of data included tide level, precipitation, and level of ground water. Their final results showed that SVM perform better than ANN. But, the data for model development is not sufficient as it is not promptly gotten since of cost restrictions and demonstrate instabilities in the developed model. [7] also applied ANN in their flood prediction model for a Mediterranean agro-watershed data. The data sets included precipitation, river flow discharge, evapotranspiration, wind speed, humidity, temperature and solar radiation. The result from the implementation of ANN gave finest measurable values of hourly flow of observation and prediction based on the average values from three different regional climate area. Unfortunately, ANN cannot re-enact uncommon occasions with high precision. This is due to the [7] having trouble to extrapolate ANN well past their limits of training. Again, [8] used ANN to predict flood by assessing the climate change impact on river runoff. They incorporated ANN with statistical model and packaged it as a hydrological model. The data set included the daily time series data for rainfall and runoff temperature. However, ANN is not able to produce the same pattern for series of daily and annual rainfall.

Another group of researchers [9]-[12] compared SVM with other statistical models and hybrid algorithms. [9] studied the performance of SVM with Support Vector Regression (SVR) model by studying rainfall pattern. SVR produced slightly better results compared to SVM. However, both models have an issue of over predicting the amount of run-off data and some limitations regarding model calibration. Statistical downscaling model has been compared with SVM in [10] by using precipitation data. Their results showed that SVR produced less accurate values compared to the SVM. [11] developed a flood prediction model wavelet-SVM and compared with SVR for the stream flow of data type. Their result could be improved further by separating the inputs or transforming the wavelet robustly. A hybrid of SVR with Ensemble Empirical Mode Decomposition (EEMD) is proposed in [12] to predict monthly rainfall. However, the limitation of EEMD somehow reduce the accuracy of monthly rainfall forecasting of SVR-EEMD.

Finally, [13]-[17] proposed and compared DT with regression model, SVR, fuzzy model, logistic models, Bayes model, multivariate, and SVM in the flood prediction domain. [15] showed that fuzzy DT performed well compared to crisp DT. Fuzzy approach have more potential in analyzing flood patterns which can be used to present the process of flood. The accuracy values from crisp DT were too sensitive to noise. [16] concluded that DT model has the highest accuracy compared to Bayes and logistic models. However, their alternative DT model does not think about the on-stationary signals or changes of temporal.

There is a limited works that compared the performance of ANN, SVM and DT together in a same domain particularly in flood prediction with the same data sets. It is worth to explore these three machine learning algorithms and do the evaluation based on their accuracy values. Thus, to compare the accuracy of each machine learning algorithms in flood prediction, ANN, SVM and DT were chosen and compared empirically for this research project.

3. METHODOLOGY

Machine learning approach plays a major role in prediction. It helps in prediction analysis as it is a quick method in eliminating data that is not related and will speed up or increase the process of analyzing situations. The translation and examination of incorporated and enormous information is not applicable and it is insufficient if it is done in a traditional way.

For this research project, flood disaster management will be focused. Three machine learning models which are ANN, SVM and DT are suitable for predicting flood as they are well-known in prediction analysis.

3.1 Artificial Neural Network

ANN standouts amongst the most well-known modeling in flood prediction. In comparison with traditional statistical models, ANN approach has a greater accuracy in prediction [18].

One of the reason is because it is an insightful procedure that copies the biological nervous systems. It works like a human brain. It consists of a system of neurons which are connected by synapses or in other word, a collection of connected units or nodes [19].

All in all, ANN modeling comprises of input layer, hidden layer and output layer which are the three fundamental layers as shown in Figure 1. The neurons' number can be manipulated. The neurons in hidden layer is important as it sends the information to next operation. The quantity of neurons assumes a noteworthy job in forecast since it influences the network performance. For instance, if it is too few neuron used, the outcome will be under-fitting while if there are too many neurons, the network will be over-fitting [19].

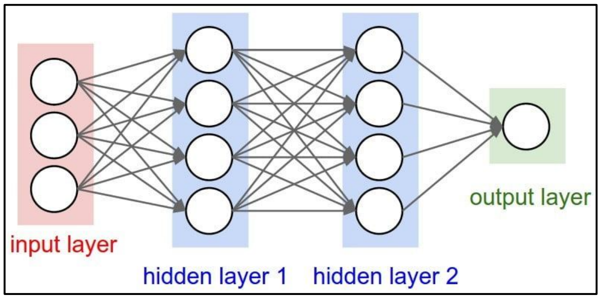

Figure 1: Structure of ANN model

3.2 Support Vector Machine

SVM is a supervised machine learning which is applied mostly in flood modeling. It is utilized to foresee a measure of time depends on the training from past and historical data. Apart from ANN as one of the machine learning method, it is stated that SVM has been an alternative method for flood prediction as it is quite known among hydrologists [2]. SVMs algorithm were applied in most of the cases involve in flood prediction which results in better performance and excellent in generalization ability. It is mostly applied in classification and regression problems.

The application of SVM diminishes over-fitting and anticipated mistakes of learning machines by receiving the hypothesis of auxiliary hazard minimization. After deciding the support vectors and appropriate kernel filters, in numerous cases SVMs may be more productive than ANN methods. The network architecture of SVM is shown in Figure 2.

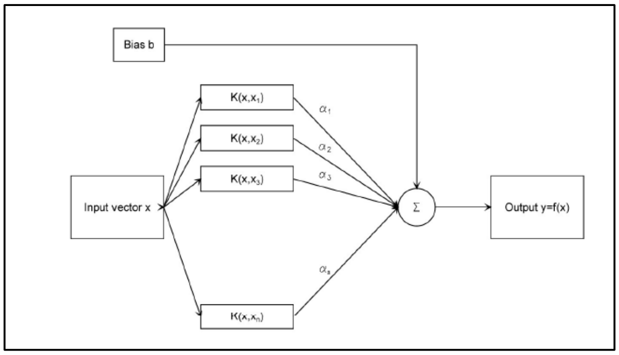

Figure 2: Network architecture of SVM

3.3 Decision Tree

DT contributes in predictive modeling especially in flood simulation. It works by utilizing a tree of choices from branches to the targeted estimations of leaves as shown in Figure 3. According to [2], DTs has been one of fast algorithms as they are well known in flood modeling and flood prediction. Decision tree are applied in classification and regression.

According to [20], one of the algorithms in predicting continuous dependent variables is regression trees which is known as tree-building algorithms. It works by separating the indicator information into subdivision of smaller districts to get estimated structure of nonlinear regression. The datasets are split and partitioned into two-sub spaces which can improve the prediction accuracy. As a result, decision tree algorithms outperformed existing models in selecting more critical harm impacting factors and in inferring multi-variate flood damage models. It is demonstrated that decision tree models are an alternative to traditional model.

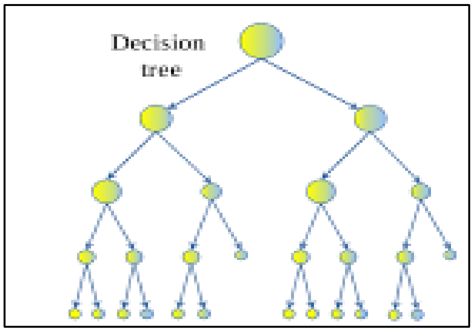

Figure 3: DT model

4. RESULTS

Dataset 1 is collected and retrieved from Department of Irrigation and Drainage (DID), Penang, Malaysia which consist of average of heavy rainfall (mm), duration of flood or rainfall (hour), flood depth (m) and flood area (km2) based on the flood annual report from the year 2017 in Penang. Another set of datasets are based on flood reports in Metro Manila, Phillippines [21] consist of specific coordinates such as latitude and longitude, average precipitation, land elevation and flood.

Table 1: Comparison of ANN, SVM, and DT for Test Size 0.2, 0.4, 0.6 and 0.8 using Dataset 1 (DID Dataset)

		Test Size				**Mean**
		0.2	0.4	0.6	0.8	
ANN	MAE	0.0032	0.0014	0.0009	0.0010	0.0017
	MSE	0.0004	0.0002	0.0009	0.0006	0.0006
	RMSE	0.0196	0.0126	0.0096	0.0079	0.0125
	Accuracy	**98.08%**	**98.08%**	**98.08%**	**98.08%**	**98.08%**
SVM	MAE	0.1036	0.1018	0.1024	0.1137	0.1054
	MSE	0.0120	0.0110	0.0113	0.0202	0.0137
	RMSE	0.1095	0.1049	0.1064	0.1421	0.1158
	Accuracy	**54.97%**	**18.58%**	**16.23%**	**0.36%**	**22.54%**
DT	MAE	0.0041	0.0021	0.0018	0.0051	0.0033
	MSE	0.0007	0.0003	0.0004	0.0015	0.0008
	RMSE	0.0262	0.0185	0.0191	0.0393	0.0258
	Accuracy	87.22%	87.36%	85.54%	54.63%	78.69%

Based on the results from Table 1 of Dataset 1, ANN has the highest accuracy with average accuracy of 98.08%, followed by DT, second highest with average accuracy of 78.69% and SVM with the least accuracy with average accuracy of 22.36%.

The results for Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RSME) for test size of 0.2,0.4,0.6 and 0.8 is shown in Table 1. For ANN, the value of MAE for test size of 0.2, 0.4 and 0.6 decreases with 0.0032, 0.0014 and 0.0009 respectively then increases when test size of 0.8 with 0.0010. Mean Squared Error (MSE) for test size of 0.2 and 0.4 decreases with 0.0004 and 0.0002 respectively then increases when test size of 0.6 with 0.0009 and then decreases to 0.0006 when test size of 0.8. Finally, Root Mean Squared Error (RSME) for test size of 0.2, 0.4, 0.6 and 0.8 decreases with 0.0196, 0.0126, 0.0096 and 0.0079 respectively.

The results for Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RSME) for test size of 0.2,0.4,0.6 and 0.8 is shown in Table 2. For ANN, the value of MAE for test size of 0.2 and 0.4 increases with 0.3357 to 0.3928 and decreases with 0.3180 when test size of 0.6 then increases with 0.3391 when test size of 0.8. Mean Squared Error (MSE) for test size of 0.2 and 0.4 increases with 0.0004 and 0.01793 then decreases when test size of 0.6 with 0.1742. The value of test size of 0.8 remains the same with 0.1742. Finally, Root Mean Squared Error (RSME) for test size of 0.2 and 0.4 increases from 0.4083 to 0.4234 and decreases to 0.4174 when test size of 0.6 and remains the same when test size of 0.8.

Table 2: Comparison of ANN, SVM, and DT for Test Size 0.2, 0.4,

		Test Size				**Mean**
		0.2	0.4	0.6	0.8	
ANN	MAE	0.3357	0.3928	0.3180	0.3391	0.3464
	MSE	0.1667	0.1793	0.1742	0.1742	0.1736
	RMSE	0.4083	0.4234	0.4174	0.4174	0.4167
	Accuracy	**77.03 %**	**77.21 %**	**77.07 %**	**77.07 %**	**77.10 %**
SVM	MAE	0.2641	0.2567	0.2545	0.2486	0.256
	MSE	0.1457	0.1355	0.1257	0.1401	0.1368
	RMSE	0.3817	0.3680	0.3545	0.3742	0.3696
	Accuracy	**23.01 %**	**25.74 %**	**29.91 %**	**21.61 %**	**25.07 %**
DT	MAE	0.2536	0.2678	0.2873	0.3016	0.2776
	MSE	0.2536	0.2678	0.2873	0.3016	0.2776
	RMSE	0.5035	0.5175	0.5360	0.5492	0.5266
	Accuracy	74.64%	73.22%	71.27%	69.84%	72.25%

0.6 and 0.8 using Dataset 2 (Public Dataset)

5. CONCLUSION

Each of the machine learning models involves different algorithms. Among all machine learning models, Artificial Neural Network (ANN) is said to be the foremost known learning algorithms in modeling flood prediction. In comparison to traditional statistical models, the ANN approach was used for prediction with greater accuracy as it derives the meaning from past and historical data. Support Vector Machine (SVM) is greatly popular in flood modelling which today known as robust and efficient machine learning algorithm for flood prediction. Decision Trees (DT) are classified as fast algorithms, they became very popular in ensemble forms to model and predict floods.

The performance accuracy of Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) are evaluated and compared:

- Based on the results, for Dataset 1, ANN has the highest accuracy with average accuracy of 98.08%, followed by DT, second highest with average accuracy of 78.69% and

SVM with the least accuracy with average accuracy of 22.36%.

- On the other hand, based on Dataset 2, ANN has the highest accuracy with average accuracy of 77.10%, followed by DT, second highest with average accuracy of 72.24% and SVM with the least accuracy with average accuracy of 25.07%.

Also, the relationship between test size setting and performance accuracy for each machine learning is concluded below:

- Both datasets, Dataset 1 and Dataset 2 are tested with various test size with 0.20, 0.40, 0.60 and 0.80 for test data and 0.80, 0.60, 0.40, 0.20 for train data respectively by applying the three machine learning algorithms.
- Based on the results, the value of performance accuracy of Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) for Dataset 1 and Dataset 2 decreases as test size increases.
- According to [22], the test size affects the percentage of performance accuracy of Support Vector Machine (SVM) and Decision Tree (DT) as the larger the number for train data, it will produce higher reading of accuracy and reliably conveys a much superior and steady results in prediction.

Based on the findings, it can be concluded that machine learning is one of the techniques that is successful as it is faster in prediction analysis. Flood prediction also contributes in preventing damage infrastructure and to dodge misfortune of lives.

ACKNOWLEDGEMENT

Research experiments reported were pursued under the Bridging Grant by Universiti Sains Malaysia [304.PKOMP.6316019], Research University Grant by Universiti Sains Malaysia [1001.PKOMP.8014073] and Fundamental Research Grant Scheme (FRGS) by Ministry of Education Malaysia [203.PKOMP.6711534] and [203.PKOMP.6711713]. Real datasets provided by Department of Irrigation and Drainage (DID), Penang, Malaysia.

REFERENCES

[1] P. Victor. **Flood Control in Southeast Asia**, 2017. [Online]. Available: https://theaseanpost.com/article/flood-control-southeast-asia. [Accessed 12 November 2018].

[2] P. O. K.-w. C. Amir Mosavi. **Flood Prediction Using Machine Learning, Literature Review**, pp. 5-8, 2018.

[3] M. Arslan, A. Roxin, C. Cruz and D. Ginhac. **A Review on Applications of Big Data for Disaster Management**, HAL archives-ouvertes.fr, 2017. https://doi.org/10.1109/SITIS.2017.67

[4] H. T. A. Y. S. L. M.Rezaeianzadeh. **Flood flow forecasting using ANN, ANFIS and regression models**, pp. 25-37, 2013. https://doi.org/10.1007/s00521-013-1443-6

[5] B. B. J. Q. J. A. A. P. Manish Kumar Goyal. **Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic and ANFIS**, *Expert Systems with Applications*, pp. 5267-5276, 2014. https://doi.org/10.1016/j.eswa.2014.02.047

[6] S.-C. J. Y. H. G.-O. B. K.-K. L. Heesung Yoon. **A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer**, *Journal of Hydrology*, pp. 128-138, 2011.

[7] Z. D. G. P. K. Nektarios N. Kourgialas. **Statistical Analysis and ANN modeling for predicting hydrological extremes under climate change scenarios: The example of a small Mediterranean agro-watershed**, *Journal of Environment Management*, pp. 86-101, 2015.

[8] S. S. S. H. M. A. M. N. H. Zulkarnain Hassan. **Suitability of ANN applied as a hydrological model coupled with statistical downscaling model: a case study in the northern area of Peninsular Malaysia**, pp. 463-477, 2015. https://doi.org/10.1007/s12665-015-4054-y

[9] R. G. G. d. M. Francesco Granata. **Support Vector Regression for Rainfall-Runoff Modeling in Urban Drainage: A Comparison with the EPA's Storm Water Management Model**, *Water*, pp. 1-13, 2016.

[10] P.-S. Y. Y.-H. T. Shien-Tsung Chen. **Statistical downscaling of daily precipitation using support vector machines and multivariate analysis**, *Journal of Hydrology*, pp. 13-22, 2010.

[11] M. C. Ozgur Kisi. **A wavelet-support vector machine conjuction model for monthly streamflow forecasting**, *Journal of Hydrology*, pp. 132-140, 2011.

[12] W. L. X. X. Y. Z. W. C. T. Y. Qi Ouyang. **Monthly Rainfall Forecasting Using EEMD-SVR Based on Phase-Space Reconstruction**, *Water Resource Management*, pp. 2311-2325, 2016. https://doi.org/10.1007/s11269-016-1288-8

[13] H. K. U. B.Merz. **Multi-variate flood damage assessment: a tree-based data-mining approach,** *Natural Hazards and Earth System Sciences*, vol. 13, pp. 53-64, 2013.

[14] S. Z. A. M. B. S. N. t. H. M. Sh. Sahraei. **Daily discharge forecasting using least square support vector regression and regression tree**, pp. 410-422, 2015.

[15] D. V. J. S. Anna E. Sikorska. **Flood-type classification in mountainous catchments using crisp and fuzzy decision trees**, *Water Resources Research*, pp. 7959-7976, 2015.

[16] K. C. A. S. H. S. I. R. I. P. D. T. B. Khabat Khosravi. Binh Thai Pham. **A comparative assessment of**

decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northen Iran, *Science of the Total Environment*, pp. 744-755, 2018. https://doi.org/10.1016/j.scitotenv.2018.01.266

[17] E. M. M. G. J. A. F. S.-H. A. M. Bahram Choubin. **An ensemble prediction of flood susceptibility using multivariate discriminat analysis, classification and regression trees and support vector machines**, *Science of Total Environment,* pp. 2087-2096, 2019. https://doi.org/10.1016/j.scitotenv.2018.10.064

[18] M. Yu, C. Yang and Y. Li. **Big Data in Natural Disaster Management: A Review**, Geosciences, pp. 1-26, 2018.

[19] M. M. A. J. M. M. H. S. R. H. A. Khoo Chun Keong. **Artificial Neural Network Flood Prediction for Sungai Isap Residence**, pp. 1-6, 2016

[20] H. K. U. B.Merz. **Multi-variate flood damage assessment: a tree-based data-mining approach**, Natural Hazards and Earth System Sciences, vol. 13, pp. 53-64, 2013. https://doi.org/10.5194/nhess-13-53-2013

[21] G. X. **Kaggle.com**, 2017. [Online]. Available: https://www.kaggle.com/giologicx/aegisdataset. [Accessed April 2019].

[22] R.-A. H. Q. a. H.-K. A.R. Ajiboye. **Evaluating the effect of dataset size on predictive model using supervised learning technique**, *International Journal of Software Engineering & Computer Science,* vol. 1, pp. 75-84, 2015. https://doi.org/10.15282/ijsecs.1.2015.6.0006