Comparative Analysis of Machine Learning Algorithms for Rainfall Prediction in Kuantan, Pahang, Malaysia.

Seri Liyana Ezamzuri, Sarah 'Atifah Saruchi
Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia

Ammar A.M. Al-Talib
UCSI University, 56000 Kuala Lumpur, Malaysia

Email: miz24007@adab.umpsa.edu.my, sarahatifah@umpsa.edu.my, ammart@ucsiuniversity.edu.my

Abstract

This study compares the performance and accuracy of four ML algorithms which are Support Vector Regressor (SVR), Artificial Neural Network (ANN), Random Forest Regressor (RFR), and Linear Regression (LR) in the rainfall prediction application. All four methods employ the same input parameters which are temperature (°c), dew point (°c), humidity (%), wind speed (Kph) and pressure (Hg). Meanwhile the output parameter is set to be the rainfall (mm) which indicates the precipitation in Kuantan, Pahang, Malaysia. The analysis shows that the SVR consistently outperforms the other machine learning algorithms, achieving the lowest Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Keywords: Machine Learning (ML), Support Vector Regressor (SVR), Artificial Neural Network (ANN), Random Forest Regressor (RFR), Linear Regressor (LR), Rainfall prediction.

1. Introduction

Heavy rainfall affects open water recreational activities due to safety and precaution for natural disasters. Floods, and water surges also known as headwater incidents commonly happen in Malaysia cause from heavy pour which led to increase volume of water and velocity of water current.

Recently, there was one case reported due to a water surge incident. Three Jabatan Kerja Raya (JKR) officers were lost in a water surge incident while participating in water rafting activities in Sungai Jahang, Gopeng, Perak [1]. This evidently shows heavy downfall led to water surge, and endangered human lives and natural ecosystem.

On top of that, using advanced technology that we have nowadays, which is Artificial Intelligence (Ai), specifically Machine Learning (ML) to find out the best algorithms for Rainfall Prediction possible outcomes by leveraging patterns in historical data. Hence, this study is to find the best ML algorithms with finest predictions accuracy value.

Due to the proximity to the ocean, abundance of rivers, and location on the Malaysian peninsula, Kuantan was selected as the study's location. The dataset used in this study was sourced from Weather and Climate which provides comprehensive historical weather data for various global regions, including Kuantan, Pahang, Malaysia. The dataset includes key meteorological features such as temperature, humidity, wind speed, pressure, and precipitation, covering the period from 2018 to 2020.

2. Methodology

Various data and models are needed to make a prediction. Generally, we use classification and regression algorithms for time series algorithms [2]. Four ML algorithms have been developed in this study to find the lowest value of MSE and MAE to ensure that algorithm is the most compatible compared others. Other than that, we find the lowest accuracy value in every ML algorithm to support the conclusion in this study.

The methodology for rainfall (precipitation) prediction in Kuantan, Pahang, Malaysia studies by four machine learning algorithms- Artificial Neural Networks (ANN) [3], Support Vector Regression (SVR) [4], Random Forest Regressor (RFR) [5], and Linear Regression [6]. Details of the data collection, data preparation, feature engineering, model training, evaluation process, and calculate correlation coefficient are explained below.

2.1 Data Collection

A detailed study by [7] analyse various parameters of rainfall prediction and figure out each parameter in meteorological features. Historical weather data for rainfall prediction was obtained from Weather and Climate. Fig.1 shows raw data for all parameters involved in this study collected in Kuantan, Pahang, Malaysia from year 2018 to 2020. The dataset included

meteorological features such as Temperature (℃), Dew Point (℃), Humidity (%), Wind Speed (Kph), Pressure (Hg), and Precipitation (mm). In Fig.2, data for precipitation were separated from other parameters to observe the trend closely over the year 2018 until 2020. The data were manually copied and separated by monthly data into datasheets as database.

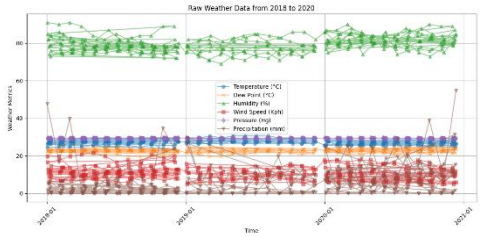


Fig. *1*. Raw Data for all parameters in 2018 until 2020

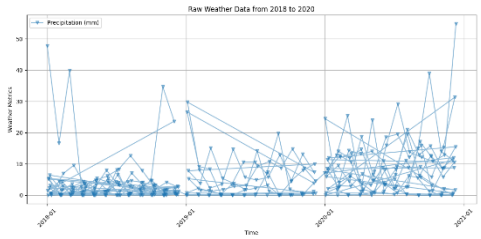


Fig. *2*. Raw data for precipitation in 2018 until 2020

2.2 Preprocessing Steps:

The implementation of ML was carried out using open-source software, Visual Studio Code as platform to code with support for operations development such as debugging, running task and control. The programming language used in this study is Python, which can be used to develop Ai. [8] said, Python language is commonly used to develop Ai due to its simple language, and it comes with built-in libraries for Ai projects usage such as NumPy, SciPy, matplotlib and some other important libraries.

Anaconda distribution tool kit was installed as computational environment framework also known as Integrated Development Environments (IDEs). Anaconda is functioned to perform Python and ML on same project and keep libraries always up to date in a single process. [9] said the most important libraries and IDEs in applications are included in Anaconda with a simple and well supported Graphical user interface (GUI). Nonetheless, Anaconda is an option for user to perform on Ai project.

Other than that, another python library, pandas were utilized are the first thing need to import right after start code a new project. Pandas library is a tool and structures specifically to perform data analysis, data manipulation and preprocessing due to its efficient handling of structured data [10]. Despite, pandas were chosen as statistical computing environments for this project.

The datasets are manually filled into .csv format and were arranged in tabular format to make it structured. There are few parameters measured which is temperature, humidity, wind speed, pressure, and precipitation. Load datasets from 2018 to 2020 into file path and ensure there is no null value or improperly formatted in the file. NumPy arrays were used to hold data into a structured format.

2.2.1 Support Vector Regression

Support Vector Regression (SVR) is one of ML algorithms which use to counter regression problem as it minimizes the prediction error. SVR have ability to convey non-linear models between precipitation and meteorological parameters.

Radial Basis Function (RBF) kernel was used since it is well-applied to non-linear models. [11] wrote, RBF kernel is a well-known kernel-based classifier since it requires a properly tuned parameter. In this study, we use hyperparameters tuning C=100 to control model complexity and achieving low error, gamma=0.1 as influence individual data points, and margin of tolerance for error prediction, epsilon=0.1. The meaning of hyperparameter is as Table 1.

Table 1. SVR Hyperparameter

Table 1 C	100	Regularization parameter
epsilon	0.1	Tube width
kernel	RBF	Type of kernel which is radial basis function (RBF)
gamma	scale	Kernel coefficient

Each ML algorithms in this study were evaluated the performance of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and accuracy of prediction within $\pm 10mm$ precipitation. Figure 4 shows value of MAE is 3.31, RMSE is 5.11, and Accuracy is 94.22% for SVR algorithms.

2.2.2 Artificial Neural Network (ANN)

Among ML algorithms, ANN is a versatile algorithm which uses interconnected nodes from input layer, hidden layer, and output layer to solve problems by mimicking structure and function of human brain [12]. The input layer contains information or parameters we get from the data to process and analyze into the next layer, which is a hidden layer. The hidden layer will take data from the input layer, and each layer analyzes the data before processing to the next layer, which is the output layer. The output layer is a combination of data from input layer, and hidden layer that have been processed and came out with single result, or more than one for multi-class classification problem.

Three hidden layer neural network in this study. The first hidden layer is an input layer containing 64 neurons with ReLu activation. The second hidden layer contains

32 neurons with ReLu activation. The third hidden layer contains 16 neurons with ReLu activation. Output Layer which generates output value and comes out with 1 neuron to predict a single output value and for this study, the output is precipitation. The meaning of hyperparameters for ANN are as Table 2.

Table 2. Hyperparameter for ANN

Hidden layer size	(64,32,16,1)	Number of neurons in
Activation	'relu'	Activation function for hidden layer
Solver	'adam'	Optimization solver
Epochs	100	Model will be train for 100 times
Batch size	16	Process 16 samples at a time
Validation split	0.2	20% of data will be use as validation data, while remaining 80% will be used to train the model

Rectified Linear Unit (ReLu) is an activation function that is applied on neurons layer with computational calculations in Python using NumPy libraries, and TensorFlow by importing dense in ANN so it will automatically apply ReLu formula into code.

$$f(x) = \max(0, x)$$

This formula will be applied into code. If the input value of x is positive, the value remains unchanged. Meanwhile if the x is negative, the value will become 0.

To evaluate the performance of ANN, we performed RMSE, MAE, and accuracy evaluation. In the figure below shows value of MAE is 3.81, RMSE is 6.03, and accuracy is 93.62% for $\pm 10mm$ of precipitation prediction.

2.2.3 Random Forest Regression (RFR)

Random Forest Regressor is an ML algorithm that constructs combination of trees predictors, and each tree determines the values of a sampled randomly consisting of tree predictors. [13] said, generalization error of forest tree classifiers depends on the number of trees in forest, and the correlation between them. The higher the number of trees in forest, the lower the error rate of generalization for forest.

RFR was found capable of handling big sets of data and handling non-linear connections. Due to this, RFR is an option for this study as in this study, we involved complicated meteorological datasets where variables such as temperature, wind speed, precipitation, and humidity show complex dependencies.

For this research, the hyperparameter of RFR is set to value 100 for estimators, and 42 random state value. Maximum depth of trees was tuned to ensure sufficient range among the decision trees. Refer to the figure below for the value of hyperparameter RFR tuning. The meaning of hyperparameters for RFR are as Table 3.

Table 3. Hyperparameter for RFR

N estimators	100	Number of trees in the forest
Random state	42	Set random seed, ensuring results are reproducible

However, this study we identified that RFR slightly underperformed compared to ANN and SVR due to its accuracy of precipitation prediction within $\pm 10mm$ is 90.58% compared to ANN and SVR which have a slightly higher value of accuracy. Additionally, refer to figure below RFR's mean absolute error (MAE) value is 3.88 and root mean squared error (RMSE) value is 5.81 which is a competitive value in between SVR, ANN, and RFR. Overall, we can say RFR is a potential, but SVR was better for this type of dataset, and prediction for this study.

2.2.4 Linear Regression

Linear regression (LR) algorithms are commonly used in ML for regression tasks. [14] study, the relationship between a dependent variable and independent variable describes the fitting of linear equation to the observed data. The general equation of LR model and its meaning are as Table 4.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \in$$

Table 4. The equation of linear regression

y	predicted value
$x_1, x_2, \ldots, x_n$	input
β_0	Intercept, representing value of y when x_i are zero
$\beta_1, \beta_2, \ldots \beta_n$	Weights on target variable
$\in$	Error

LR is one of the methods applied in this research to predict rainfall in Kuantan based on historical meteorological data that seems to match this ML algorithm. The dataset wat split into training data to estimate the coefficients (β_i) and testing to minimize the sum of squared between predicted and actual values.

Root mean squared error (RMSE) and mean absolute error (MAE) were evaluated in this study. However, after training it only shows baseline results for the predictions which is 83.89%. This study shows that LR algorithms disputed to train complex dataset which occurred to high RMSE and MAE values compared to other algorithms, SVR, ANN, and RFR.

3. Result and Analysis

This article wants to show a comparative analysis among four algorithms tested on meteorological dataset to predict precipitation in Kuantan, Pahang, Malaysia. Four ML algorithms consist of Artificial Neural Network (ANN), Support Vector Regression (SVR), Random Forest Regression (RFR), and Linear Regression (LR).

The algorithms were evaluated based on Root Mean Absolute Error (RMSE), Mean Squared Error (MSE), and accuracy within $\pm 10mm$ of precipitation. As a result, it shows patterns of predictions and actual data of rainfall, and competitive values of accuracy.

The following figure shows prediction vs. actual values of precipitation for all ML algorithms tested in this study. Fig. 3 below shows the graph plotted from Support Vector Regression (SVR) prediction algorithm, where the Precipitation as a parameter on Y-axis. While X-axis represents time which is three years of data from 2018 until 2020.

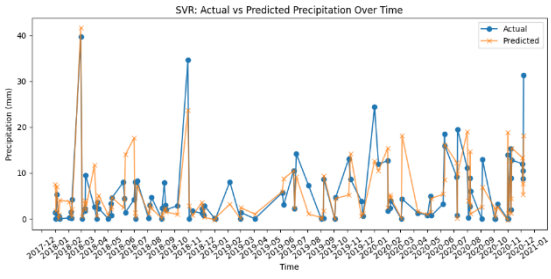


Fig. 3. Prediction data using SVR

To observe the differences in predictions, we use another algorithm to compare. In Fig.4 Artificial Neural Network (ANN) were used to predict Precipitation (mm) over Time from year 2018 until 2020.

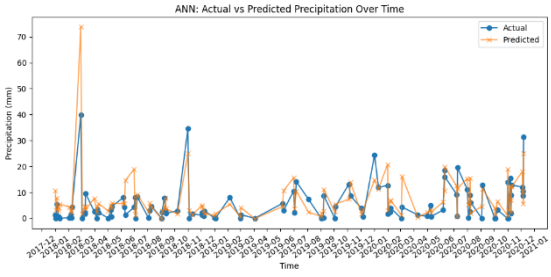


Fig. 4. Prediction data using ANN

Fig.5 shows the prediction and actual data for Precipitation (mm) over Time using Random Forest Regression (RFR) ML algorithm.

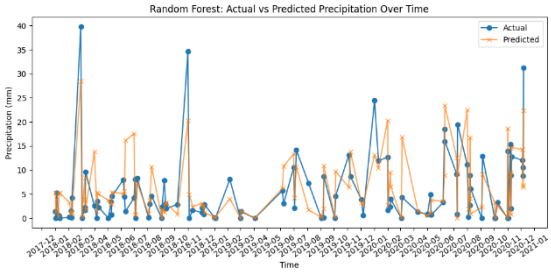


Fig. 5. Prediction data using RFR

In addition, we applied Linear Regression (LR) algorithm to compare and evaluate the prediction outcomes, and the result is as shown in Fig. 6.

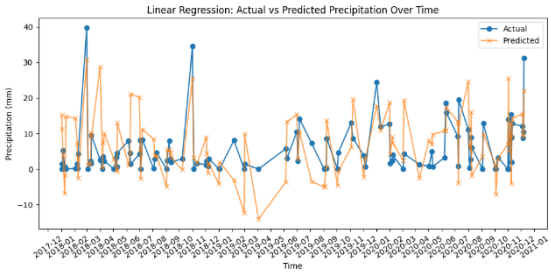


Fig. 6. Prediction data using LR

By comparing the results of all four algorithms and plotting into a graph as shown in Fig.7 will be able to see the variations, accuracy and highlight the differences between predictions and actual data.

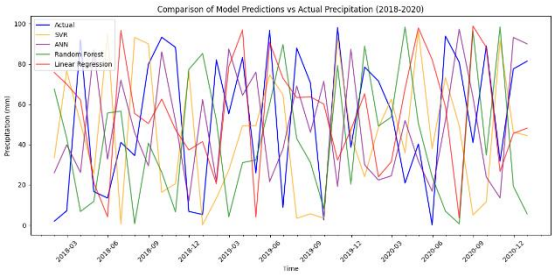


Fig. 7. Comparison of actual and predicted precipitation values from four algorithms

Table 5 Comparison results of prediction from four ML algorithms

ML Algorithms	RMSE	MAE	R^2 coefficient of determination	Accuracy of prediction 10mm (%)
SVR	5.11	3.31	0.53	94.22
ANN	6.03	3.81	0.34	93.62
RFR	5.81	3.88	0.39	90.58
LR	7.50	5.83	-0.03	83.89

The result and analysis in Table 5 show that simple ML algorithms like linear regression (LR) attempted to accurately predict precipitation, likely due to complex

dataset which need more advance ML algorithms to perform. SVR performs as the most accurate and reliable algorithms, outstripping ANN, RFR, and LR models in all metrics.

4. Conclusion

This study reveals that advanced machine learning algorithms such as SVR and ANN are able to train and predict more extensive datasets and capture non-linear relationships. Comparing both competitive result of SVR, and ANN, SVR produce more outstanding result from all metrics that have been test, which is RMSE, MAE, R^2 and Accuracy. This highlights its ability to handle complex datasets and non-linearities. These findings present the importance of selecting proper algorithms for complex predictive datasets and optimizing hyperparameters to enhance the result of model's predictive capabilities.

Acknowledgements

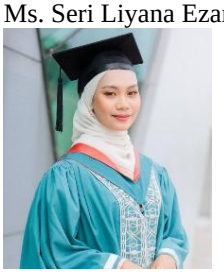
This study was funded by Yayasan Pahang through Universiti Malaysia Pahang Al-Sultan Abdullah Agency Grant RDU240705.

References

1. N. Trisha, "Sg Jahang tragedy: Victim's brother didn't expect phone call was to be their last," *The Star*, Ipoh, pp. 1–1, Nov. 16, 2024.
2. J. Faouzi, "Time Series Classification: A Review of Algorithms and Implementations," 2024. doi: 10.5772/intechopen.1004810.
3. S. Walczak and N. Cerpa, "Artificial Neural Networks," in *Encyclopedia of Physical Science and Technology (Third Edition)*, R. A. Meyers, Ed., New York: Academic Press, 2003, pp. 631–645. doi: https://doi.org/10.1016/B0-12-227410-5/00837-1.
4. C. Gambella, B. Ghaddar, and J. Naoum-Sawaya, "Optimization problems for machine learning: A survey," *Eur J Oper Res*, vol. 290, no. 3, pp. 807–828, 2021, doi: https://doi.org/10.1016/j.ejor.2020.08.045.
5. M. Diamantopoulou, "Simulation of over-bark tree bole diameters, through the RFr (Random Forest Regression) algorithm," *Folia Oecologica*, vol. 49, pp. 93–101, Jul. 2022, doi: 10.2478/foecol-2022-0010.
6. K. Kumari and S. Yadav, "Linear regression analysis study," *Journal of the Practice of Cardiovascular Sciences*, vol. 4, p. 33, Jan. 2018, doi: 10.4103/jpcs.jpcs_8_18.
7. M. S. Pathan, J. Wu, Y. H. Lee, J. Yan, and S. Dev, "Analyzing the Impact of Meteorological Parameters on Rainfall Prediction," in *2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium)*, 2021, pp. 100–101. doi: 10.23919/USNC-URSI51813.2021.9703664.
8. S. Mihajlovic, A. Kupusinac, D. Ivetic, and I. Berković, *The Use of Python in the field of Artifical Intelligence*. 2020.
9. D. Rolon-Merette, M. Ross, T. Rolon-Merette, and K. Church, "Introduction to Anaconda and Python: Installation and setup," *Quant Method Psychol*, vol. 16, pp. S3–S11, May 2020, doi: 10.20982/tqmp.16.5.S003.
10. W. Mckinney, "pandas: a Foundational Python Library for Data Analysis and Statistics," *Python High Performance Science Computer*, Jan. 2011.
11. X. Ding, J. Liu, F. Yang, and J. Cao, "Random Radial Basis Function Kernel-based Support Vector Machine," *J Franklin Inst*, vol. 358, Oct. 2021, doi: 10.1016/j.jfranklin.2021.10.005.
12. P. Yu, M. Low, and W. Zhou, "Design of experiments and regression modelling in food flavour and sensory analysis: A review," *Trends Food Sci Technol*, vol. 71, Nov. 2017, doi: 10.1016/j.tifs.2017.11.013.
13. L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
14. R. Sarmento and V. Costa, "Introduction to Linear Regression," 2017. doi: 10.4018/978-1-68318-016-6.ch006.

Authors Introduction

Ms. Seri Liyana Ezamzuri



Graduated from Bachelor's degree in Mechatronics Engineering in 2023 from the Faculty of technology manufacturing and mechatronics in Universiti Malaysia Pahang Al-Sultan Abdullah, Pahang, Malaysia. Currently a Msc. Student in Universiti Malaysia Pahang Al-Sultan Abdullah, Pahang, Malaysia.

Dr. Sarah 'Atifah Saruchi

She received her Bachelor's in Mechatronics from Nagoya University, Japan, and completed her Master's and Ph.D. at Universiti Teknologi Malaysia. She is currently a lecturer at Universiti Malaysia Pahang

Asst. Prof. Dr. Ammar Al-Talib

He received his B.Sc and M.Sc degrees in Mechanical Engineering from the University of Mosul Iraq. He has finished his Ph.D degree from UPM University, Malaysia. Currently, he is working in UCSI University.