International Journal of Engineering & Technology, 7 (3.22) (2018) 55-58

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Flood prediction of Sungai Bedup, Serian, Sarawak, Malaysia using deep learning

Abdulrazak Yahya Saleh Al-Hababi[1]*, Roselind Tei[1]

[1]FSKPM Faculty, University Malaysia Sarawak (UNIMAS), Kota Samarahan, 94300 Sarawak, Malaysia
**Corresponding author E-mail: ysahabdulrazak@unimas.my*

Abstract

This paper aims to evaluate the performance of the Long Short Term Memory (LSTM) model for flood forecasting. Seven data sets provided by the Drainage and Irrigation Department (DID) for Sungai Bedup, Serian, Sarawak, Malaysia are used for evaluating the performance of LSTM algorithm. Distinctive network was trained and tested using daily data obtained from the DID with the year range from 2014 to 2017. The performance of the algorithm was evaluated based on (Training Error Rate, Testing Error Rate, Loss, Accuracy, Validate Loss and Validate Accuracy) and compared with the Backpropagation Network (BP). Among the seven data sets, Sungai Bedup showed small testing error rate which is (0.08), followed by Bukit Matuh (0.11), Sungai Teb (0.14), Sungai Merang (0.15), Sungai Meringgu (0.12), Semuja Nonok (0.14) and lastly Sungai Busit is (0.13). Moreover, the developed model performance is evaluated by comparing with BP model. Results from this research evidently proved LSTM models is reliable to forecasting flood with the lowest testing error rate which is (0.08) and highest validate accuracy (92.61%) compared to BP with testing error rate (0.711) and validate accuracy (85.00%). Discussion is provided to prove the effectiveness of the model in forecasting flood problems.

Keywords: Artificial Neural Network (ANNs); Backpropagation (BP); Deep Learning; Flood Forecasting; Long Short Term Memory Network (LSTM)

1. Introduction

Flood is one of the natural calamities that Malaysia faces almost every year in varying degree of magnitude. Throughout Malaysia, including Sabah and Sarawak, an estimated of 9% total area of Malaysia is vulnerable to flood and approximately almost 4.82 million people are affected by flood [1]. Over the past decade, different kinds of modelling and data types emerged to forecast the flood events [2]. Excessive rainfall can cause flooding, especially in rural and urban areas, which may undergo demographic changes from time to time. [3], data from 50 years ago reveal that 41% of all natural disasters are related to severe weather conditions or water event phenomenon such as flood. The historical record of the catchments is important to display the information and investigate the time series of flash floods occurring hourly [4]. An early accurate prediction of the occurrences is considered to overcome and reduce the impacts of flood events.

One of the previous methods selected to reduce the flood is by implementing the Artificial Neural Network (ANNs) to forecast the hourly water level. ANNs have been extensively used in different kinds of research, especially for forecasting purposes [5]. ANNs learns by tracking examples; with the specific training and learning process, it can process a set of given data. ANNs can minimise the forecasting error by implementing various algorithms in order to get the best algorithm that will yield the closest result with the actual values given, and it is a non-structural countermeasure [5].

ANNs have been developed to resemble the human biological neural network. However, the difference is that a human only processes certain information at certain times, but Deep Learning, which are developed with the same concept as human neurons can process thousands of pieces of information in a much shorter time [6].

Different principles have been used to forecast floods, such as computer simulations based on the watershed demographic model, principle of hydrological, hydraulic components and groundwater flow model [7]. However, these methods only can predict certain catchment or basin based on certain water-level value. The target in this research is a big pool of historical flood data employed to predict an accurate output; the results will then be utilized to reduce the impacts of floods not only on the society but also on the environment. The remaining parts of this paper are organized as follows: methodology is presented in section 2, empirical studies are presented in section 3, section 4 discusses the results and finally section 5 concludes the paper future works.

2. Methodology

Recently, DL has attracted a growing research interest, and the method has shown certain advantages of learning [8]. DL is able to learn from the past data to solve complex problems and has been widely used in the field of forecasting. It allows the computational models that contain numerous processing layers to learn the data given with multiple levels of abstraction [9]. The results of the DL method are compared with the standards of other neural networks. This comparison is to analyse the effectiveness of the DL algorithms in this study. This research focuses on flood forecasting in the region of Bedup River Basin. The catchment is part of the Sadong Basin, and it is located 80 km away from Kuching. [10], the area of the whole Sadong Basin is about 3550 square km2 while the total length of the main river is 150 km, as shown in Figure 1. The data sets has been collected from the DID for the years of 2014 to 2017. Forecasting has been conducted on seven

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

different gauging stations in Serian Division; which are Sungai Bedup, Bukit Matuh, Semuja Nonok, Sungai Busit, Sungai Merang, Sungai Meringgu and Sungai Tep.

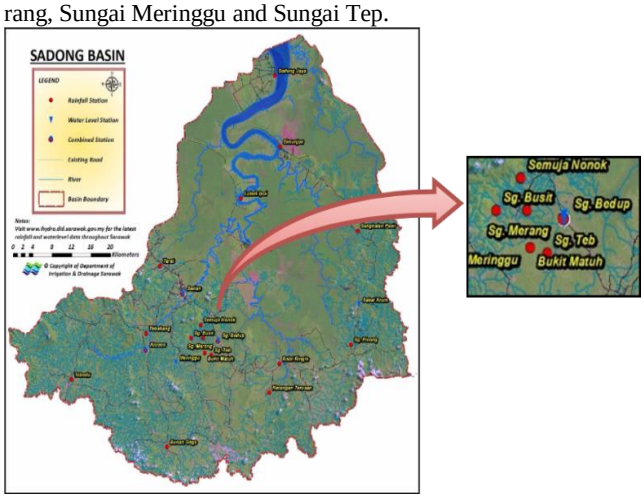

Fig. 1: Location of the Bedup River, Sarawak, Malaysia.

2.1. Long Short-Term Memory (LSTM)

LSTM architecture model is a memory cell which can maintain its state. The non-linear units of LSTM regulate the information flow in and out of the cell. Most modern studies incorporated from many of the improvements that have been made to the LSTM architecture model since its original formulation. LSTM are now applied to many learning problems which differ the significance in scale and nature from the problems that these improvements were initially tested on. A schematic of the LSTM architecture model block can be seen in Figure 2. It features three gates; input gates, forget gates, output gates, block input, a single cell of the constant error carousel, an output activation function, and peephole connections. The output of the block is recurrently connected back to the block input and all of the gates [11].

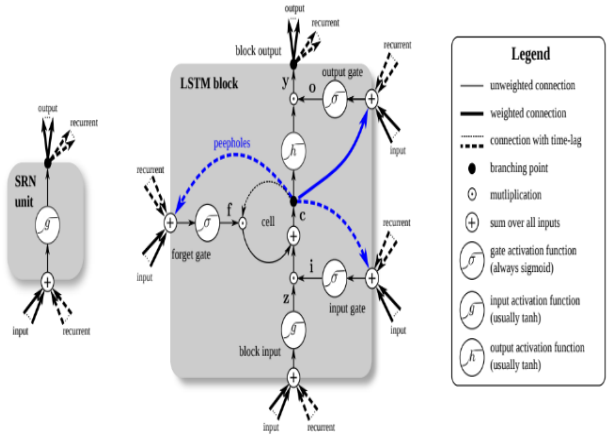

Fig. 2: Schematic of the Long Short-Term Memory Block

[12], the final weight derivatives are found by summing over the derivatives at each timestep, where *O* is the objective function used for training. *f() (frequently noted as σ(.))* is the standard logistic sigmoid function defined in Equation (1), *g()* and *h()* are the transformations of *function ()* whose range are *[-2,2]* and *[- 1,1]*.

$$\delta_j^t = \frac{\partial O}{\partial a_j^t} \tag{1}$$

The order in which the Equations (2) to Equation (6) are calculated during the forward and backward passes is important and should proceed as specified below [12]. As with standard LSTM, all states and activations are set to zero at t = 0, and all δ terms are zero at t = T + 1. LSTM will decide what information is going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer". Where ht−1 and xt, and outputs a number between 0 and 1 for each number in the cell state Ct−1.

Input Gates:

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota}\ x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1} + b_\iota^t = f(a_\iota^t) \tag{2}$$

Forget Gates:

$$a_\emptyset^t = \sum_{i=1}^{I} w_{i\emptyset}\ x_i^t + \sum_{h=1}^{H} w_{h\emptyset} b_h^{t-1} + \sum_{c=1}^{C} w_{c\emptyset} s_c^{t-1} + b_\emptyset^t = f(a_\emptyset^t) \tag{3}$$

Cells:

$$a_c^t = \sum_{i=1}^{I} w_{ic}\ x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1} + s_c^t = b_\emptyset^t s_c^{t-1} + b_\iota^t = g(a_c^t) \tag{4}$$

Output Gates:

$$a_w^t = \sum_{i=1}^{I} w_{iw}\ x_i^t + \sum_{h=1}^{H} w_{hw} b_h^{t-1} + \sum_{c=1}^{C} w_{cw} s_c^{t-1} + b_w^t = f(a_w^t) \tag{5}$$

Cell Outputs:

$$b_c^t = b_w^t h(s_c^t) \tag{6}$$

Where:

- w_{ij} the weight of the connection from unit i to unit *j*
- a_i^t the network input to some unit j at time *t*
- b_i^t the value of the same unit after the activation function has been applied
- *ι* input gate, Ø forget gate, *w* output gate
- *C* set of memory cells of the block
- s_c^t state of cell c at time *t*
- *f* the activation function of the gates, g cell input activation functions, *h* cell output activation functions
- *I* the number of inputs, *K* the number of outputs, *H* number of cells in the hidden layer

3. Empirical Studies

To evaluate the performance of the LSTM model, several experiments are conducted on seven real data set given by DID. The characteristics of the data sets are shown in Table 1, which have been described considerable in data sets characteristics, features, classes and pattern sample among the seven data sets. They are the real-world data sets given by the DID that are same with respect to the number of available samples, data sets characteristics (Multivariate), Features (2) and Pattern (2). Flood forecasting is the real-world problems that have been the studies of many researchers.

Table 1: Summary of Data Sets used in this research

Data Set	Characteristics	Features	Patterns
Sungai Bedup	Multivariate	2	2
Bukit Matuh	Multivariate	2	2
Sungai Teb	Multivariate	2	2
Sungai Merang	Multivariate	2	2
Sungai Meringgu	Multivariate	2	2
Semuja Nonok	Multivariate	2	2
Sungai Busit	Multivariate	2	2

This research has been divided the original data set into two parts, where 70% of the data in the whole dataset were used as a training set for the parameters estimation of LSTM and remaining 30% of the data as a test to evaluate the performance of the models. This

research applies the data with three and a half-year for training and six months for the testing based on the details of Table 2.

Table 2: Training and Testing Period for the data

Network	Training	Testing
2014-2017	Year 2014- January to June 2017	July – December 2017

A Python implementation of deep neural networks with the configuration from the previous section was used to train the model. The Tensorflow library was used to build and compile the LSTM model on calculated using the same computer (PC with processor Intel (R), Core (TM), I5-3337U, CPU@1.80Hz and 12 Installed memory (RAM) for the seven data sets. The proposed method was implemented through three steps: data collection, preprocessing and validation phase. It has been perceived with the predictive analysis and other performance measure (Training Error Rate, Testing Error Rate, Loss, Accuracy, Validate Loss and Validate Accuracy, respectively). The test performance through LSTM has been compared with the BP. LSTM starts by collecting, normalizing, and reading the dataset. The data set goes through three main gates which are input gate, forget gate and output gate. The selected hyper-parameter has been used in the training and validating algorithm are batch size, epoch, activation and optimization as shown in Table 3. The maximum epoch is then set to the model and the proposed method stops after the maximum number of epoch is reached. The algorithms show the predictive analysis accuracy on each data set. Figure 3 briefly describes the schematic representation of LSTM model.

Table 3: Hyper-Parameter of LSTM

	Batch Size	Epoch	Activation	Optimization
Dataset	32	200	SoftMax	Adam

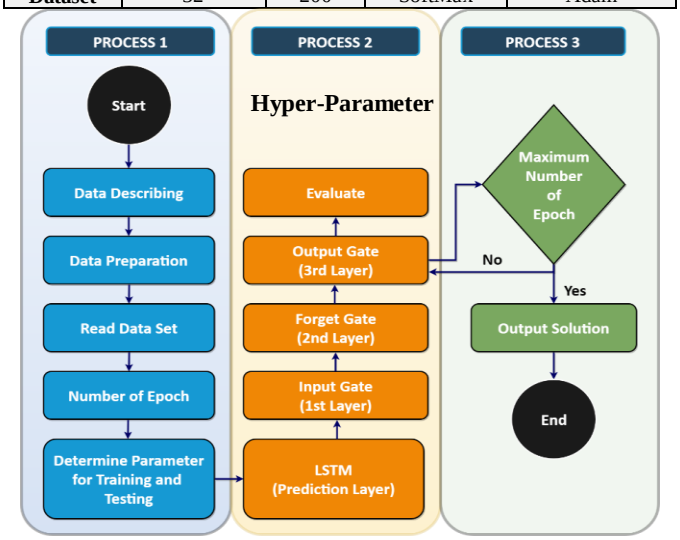

Fig. 3: Schematic representation of the proposed LSTM model

The predictive analysis used to test the accuracy of the output. The equation of predictive analysis for the dataset is given in equations (7) below. Where M is the testing part pattern, n is the class of pattern m, and $cf\ (m)$ gives the accuracy of m by the proposed method.

$$\text{Accuracy (M)} = \frac{\sum_{i=1}^{|M|} \text{evaluate } (m_i)}{|M|} \quad (7)$$

4. Results

This section presents the result of LSTM: statistical result of the seven data sets. The results of LSTM reveal the generalization of LSTM model for all seven data set. The results of all data sets involved are analyzed based on (Training Error Rate, Testing Error Rate, Loss, Accuracy, Validate Loss and Validate Accuracy). The results of the proposed method for each data set are analyzed and presented in the following subsection. In the Table 4, the best result is highlighted in the bold font. The statistical results of LSTM show that Sungai Bedup obtained the best result compared with all the data sets, respectively. In this context, Sungai Bedup contains small testing error rate **0.08** and highest validation of accuracy **92.61%**. The result of LSTM for all data sets shows high validation accuracy except Sungai Merang data set. This is due to the data set of Sungai Merang given by DID contains small parts of missing data. In this context, the missing data might result and affect the accuracy of forecasting. Moreover, the statistical result indicated that LSTM shows superior testing and validation accuracy results for all data sets. Even there is a missing data in Sungai Merang data set, LSTM is able to show highest values compared to the BP. The experiment results during the training and testing for all data sets has been presented in Table 4 and Figure 4 to Figure 5.

Table 4: Summary of LSTM results for all data sets

Data Set	Training Error rate	Testing Error rate	Accuracy	Validate Accuracy %
Sungai Bedup	0.06	**0.08**	0.8925	**92.61**
Bukit Matuh	0.10	0.11	0.8870	92.33
Sungai Teb	0.10	0.14	0.6089	88.00
Sungai Merang	0.11	0.15	0.7116	86.99
Sungai Meringgu	0.11	0.12	0.8616	90.41
Semuja Nonok	0.11	0.14	0.8925	90.00
Sungai Busit	0.11	0.13	0.6068	88.15

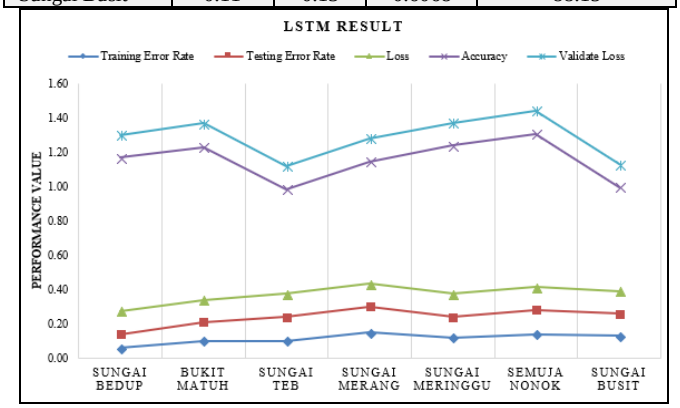

Fig. 4: Comparison of seven data sets

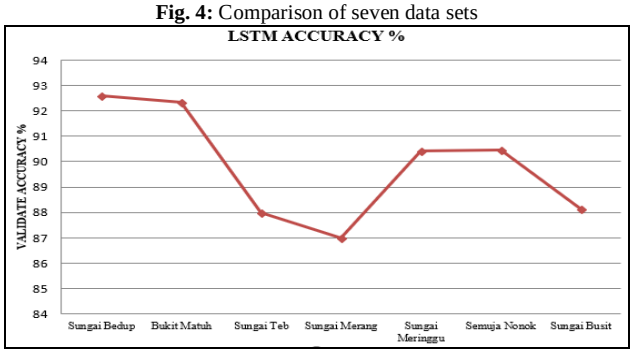

Fig. 5: Validation accuracy for all data sets

4.1. Comparative Study

This study focused on forecasting flood and it was measured by (Training Error Rate, Testing Error Rate, Loss, Accuracy, Validate Loss and Validate Accuracy, respectively). The performance of the developed models is evaluated by comparing with BP. Table 5 and Figure 6 shows a comparative analysis of the LSTM model and BP respectively. Results from this study evidently prove that LSTM models are reliable to forecast the water level at Bedup River with the lowest testing error rate which is (0.08) compared with the BP testing error rate (0.711). LSTM

achieved the highest accuracy (92.61%) compared to BP (85.00%) as shown in Figure 7. The LSTM modelling framework captures important drivers in the flood forecasting.

Table 5: Comparison of Backpropagation and LSTM

Algorithms	Training	Testing	Accuracy%
Backpropagation	1	0.711	85.00
LSTM	**0.06**	**0.08**	**92.61**

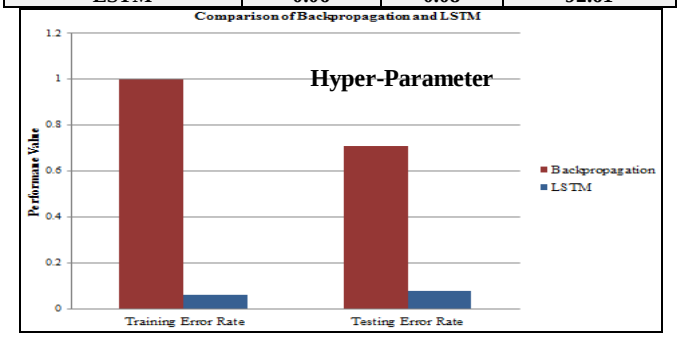

Fig. 6: Comparison of Training and Testing of BP and LSTM

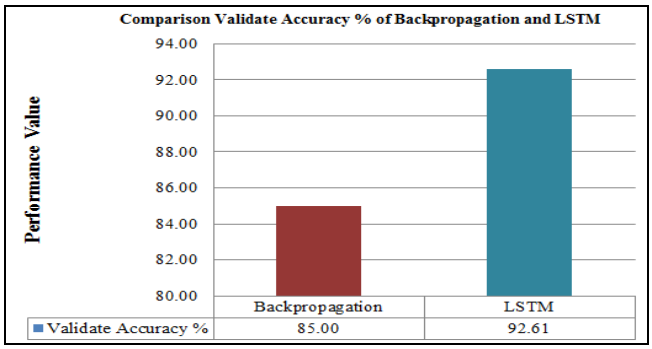

Fig. 7: Comparison Accuracy of BP and LSTM

5. Conclusion

In this research, LSTM model has been utilized for forecasting flood and compared with BP. Among the seven data sets, Sungai Bedup showed a small testing rate which is (0.08), follows by Bukit Matuh (0.11), Sungai Teb (0.14), Sungai Merang (0.15), Sungai Meringgu (0.12), and Semuja Nonok (0.14) and lastly is Sungai Busit (0.13). The performance of the developed models is evaluated by comparing them with BP model. Results from this study evidently prove that LSTM models are reliable to forecasting flood with the lowest testing error rate which is 0.08 and highest validate accuracy 92.61%. The LSTM modelling framework captures important drivers of data set demand. In the future, a combination of experience and good preferences skills will help an individual in selecting and checking the best method in the success of modelling effort. More data sets are required in the future study to enhance the result of the forecasting. The latest generation of spiking neural network will be implement and another study area will be selected to do flood forecasting on water level, Tide Level and Rainfall dataset provide by DID.

Acknowledgement

This research is supported and funded by RIMC of University Malaysia Sarawak (UNIMAS), under the special Grant Scheme (F04/SpGS/1547/2017).

References

[1] Official Website of Department of Irrigation and Drainage Sarawak (DID).[s.d.].retrieved<http://www.did.sarawak.gov.my/modules/web/pages.php?mod=webpage&sub=page&id=319>. accessed 03/Jul./18.

[2] Hapuarachchi, H. A. P.; Wang, Q. J.; Pagano, T. C. A review of advances in flash flood forecasting. Hydrological Processes, [s.l.], vol. 25, no. 18, p. 2771–2784, 2011. ISSN: 08856087, DOI 10.1002/hyp.8040.

[3] Miller, S. G. Wind, Rain, Heat: Health Risks Grow with Extreme Weather. Live Science . 2017. retrieved <https://www.livescience.com/57936-climate-change-extreme-weather-health.html>. accessed 26/Sep./17.

[4] Archer, D. R.; Parkin, G.; Fowler, H. J. Assessing long term flash flooding frequency using historical information. Hydrology Research, [s.l.], vol. 48, no. 1, p. 1–16, 2017. ISSN: 0029-1277,10.2166/nh.2016.031.

[5] Liu, F.; Xu, F.; Yang, S. A Flood Forecasting Model Based on Deep Learning Algorithm via Integrating Stacked Autoencoders with BP Neural Network. In: 2017 IEEE Third International Conference on Multimedia Big Data (BigMM). [s.l.]: IEEE, 2017. retrieved <http://ieeexplore.ieee.org/document/7966716/>. accessed 28/Sep./17. ISBN: 978-1-5090-6549-3, DOI: 10.1109/BigMM.2017.29.

[6] Maind, S. B.; Wankar, P. Research Paper on Basic of Artificial Neural Network. International Journal on Recent and Innovation Trends in Computing and Communication, [s.l.], vol. 2, no. 1, p. 96–100, 2014.

[7] Moges, M. A. et al. Suitability of Watershed Models to Predict Distributed Hydrologic Response in the Awramba Watershed in Lake Tana Basin. Land Degradation & Development, [s.l.], vol. 28, no. 4, p. 1386–1397, 2017. ISSN: 10853278, DOI: 10.1002/ldr.2608.

[8] Kruskal, J. B. et al. Big Data and Machine Learning—Strategies for Driving This Bus: A Summary of the 2016 Intersociety Summer Conference. Journal of the American College of Radiology, [s.l.], vol. 14, no. 6, p. 811–817, 2017. ISSN: 15461440, DOI: 10.1016/j.jacr.2017.02.019.

[9] Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature, [s.l.], vol. 521, no. 7553, p. 436–444, 2015. ISSN: 0028-0836, DOI: 10.1038/nature14539.

[10] Bustami, R. et al. Artificial Neural Network for Precipitation and Water Level Predictions of Bedup River Artificial Neural Network for Precipitation and Water Level Predictions of Bedup River. [s.l.], no. May 2014, 2007.

[11] Kong, Y.-L. et al. Long Short-Term Memory Neural Networks for Online Disturbance Detection in Satellite Image Time Series. Remote Sensing, [s.l.], vol. 10, no. 3, p. 452, 2017. ISSN: 2072-4292, DOI: 10.3390/rs10030452.

[12] Aichouri, i. et al. River Flow Model Using Artificial Neural Networks. Energy Procedia, [s.l.], vol. 74, p. 1007–1014, 2015. ISSN: 18766102, DOI: 10.1016/j.egypro.2015.07.832.