International Journal on Advanced Science Engineering Information Technology

Vol.6 (2016) No. 6
ISSN: 2088-5334

Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction

Suhaila Zainudin[#], Dalia Sami Jasim[#], and Azuraliza Abu Bakar

[#]*Faculty of Information Science and Technology,Universiti Kebangsaan Malaysia, UKM Bangi, 43600, Selangor, Malaysia*
E-mail: {suhaila.zainudin, azuraliza}@ukm.edu.my, daliasami99@yahoo.com

***Abstract*— Climate change prediction analyses the behaviours of weather for a specific time. Rainfall forecasting is a climate change task where specific features such as humidity and wind will be used to predict rainfall in specific locations. Rainfall prediction can be achieved using classification task under Data Mining. Different techniques lead to different performances depending on rainfall data representation including representation for long term (months) patterns and short-term (daily) patterns. Selecting an appropriate technique for a specific duration of rainfall is a challenging task. This study analyses multiple classifiers such as Naïve Bayes, Support Vector Machine, Decision Tree, Neural Network and Random Forest for rainfall prediction using Malaysian data. The dataset has been collected from multiple stations in Selangor, Malaysia. Several pre-processing tasks have been applied in order to resolve missing values and eliminating noise. The experimental results show that with small training data (10%) from 1581 instances Random Forest correctly classified 1043 instances. This is the strength of an ensemble of trees in Random Forest where a group of classifiers can jointly beat a single classifier.**

***Keywords*— rainfall prediction; data mining; classification; Random Forest; ensemble**

I. INTRODUCTION

Data Mining or Knowledge Discovery in Databases (KDD) process is used to discover new patterns from large datasets and has had a profound impact on the society by solving real-life problems [1]. Data mining aims to extract useful knowledge and represent the new knowledge to make it understandable. This knowledge can be utilized for future use [2]. Recently, a new wave of research has been conducted on time-series data mining. Time-series data mining is the process of analyzing the sequence of data points that contain successive measurements made over a time interval [3]. Several domains nowadays are relying on time-series data such as financial, stock market, climate change and others [4].

Climate change analysis analyzes the behavior of weather for a specific period of time [5]. The key characteristic behind climate change lies in the nature of its data that is captured in time point manner [6]. One of the climate change tasks is rainfall forecasting where specific features such as humidity and wind are used to predict rainfall in a specific location. Many techniques such as Support Vector Machine (SVM), Naïve Bayes (NB), Neural Network (NN) and others have been analyzed for rainfall forecasting. Most techniques tend to be supervised learning techniques. The key point behind supervised learning technique is selecting an appropriate technique with appropriate features. The performance among such techniques widely varies which leaves room for improvement by combining multiple techniques or improving present techniques.

Rainfall forecasting is a challenging task to predict factors associated with rainfall such as wind, humidity, and temperature. Basically, rainfall forecasting task is usually performed using supervised learning techniques. Since there are many different supervised learning techniques, different performances could be gained from them. In addition, rainfall data could be formed in different forms including long-terms (e.g. months) and short-terms (e.g. daily). Therefore, selecting an appropriate technique for a specific duration of rainfall is a crucial task.

Several approaches have been proposed for rainfall forecasting for many locations such as Korea, China, South Africa and others [7], [8], [9]. The current techniques for rainfall prediction including Neural Network [10], K-Nearest Neighbor and Naive Bayes [11], Support Vector Machine [12] and others. Hence, there is a need to investigate multiple techniques in order to identify the best performance in terms of rainfall prediction. In addition, there is a need to investigate new locations for rainfall forecasting such as Malaysia. Therefore, this study aims to address multiple supervised learning techniques for rainfall forecasting using Malaysian data.

This study performs a comparative analysis among several supervised learning techniques including Support

Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Neural Network (NN), and Random Forest (RF) regarding their ability to predict rainfall data. The data has been collected from multiple stations in the state of Selangor in Malaysia.

The El-Nino phenomenon has wreaked havoc with the global weather patterns including rainfall [13]. This leads to increased research efforts that addressed the rainfall prediction task. Past efforts have utilized many prediction techniques, several features/indicators, and multiple preprocessing approaches. For instance, [2] proposed a new preprocessing approach using moving average and singular spectrum analysis. Such preprocessing task will be employed on the classes of the training data in order to transform it into low, medium and high categories. Then, an Artificial Neural Network (ANN) will analyze the data in order to predict the classes on an unseen portion of data (testing). Two daily mean rainfall series from Zhenshui and Da'ninghe watersheds of China have been used as datasets for experiments.

Modular Fuzzy Inference System that aims to predict monthly rainfall data collected from the northeast region of Thailand is proposed in [14]. The hypothesis of such study lays on the uncertainty of rainfall prediction where the classes usually yield many potential instances. Fuzzy set theory has been utilized in order to estimate the membership for each input variable. Each instance will be annotated with a membership value, and then a rule-based approach was implemented in order to predict the classes of each input variable.

A multilayered Artificial Neural Network with learning by back-propagation algorithm configuration has been used to analyze data from www.indiastat.com and the IMD website [7]. The input parameters for the ANN are the average Humidity and the average Wind Speed for the 8 months in 50 years from 1960 to 2010. The output parameter is average rainfall in the 8 months of every year from 1960 to 2010. Results have shown that as the number of neurons increases in an ANN, the Mean Squared Error (MSE) decreases. In other words, more data relates to lower prediction error.

A hybrid method of feature extraction and prediction technique for predicting daily rainfall data collected from National Oceanic and Atmospheric Administration (NOAA) for more than 50 years was proposed by [12]. Basically, the features consist of humidity, pressure, temperature and wind speed. Neural Network has been used to classify the instances into low, medium and high classes based on a predefined training set.

Bayesian algorithm for rainfall prediction in India using historical data collected from the Indian Metrological Department is proposed by [15]. Six features were utilized including temperature, pressure level, mean sea level, relatively humidity, vapor pressure and wind speed. The Bayesian algorithm was trained on the data based on the mentioned features. The prediction model is observed to be more accurate if the training dataset is very large.

More recently, [11] proposed a comparative study using Regression tree (CART), naïve Bayes, K-nearest neighbor and Neural Network. The dataset of 2245 samples of New Delhi rainfall records from June to September (the annual rainfall period) from 1996 to 2014 has been used including the features mean temperature, dew point temperature, humidity, sea pressure and wind speed. Neural Network performed the best with this data with 82.1% accuracy, second best is KNN with 80.7%, Regression Tree (CART) scored 80.3% while Naive Bayes provides 78.9% accuracy.

Random Forest Ensemble Classification and Regression to improve rainfall assignment during the day, night and twilight based on cloud physical properties (remote sensing data) is proposed in [16]. The results proved that the proposed method is able to assign rainfall rates with good accuracy even on an hourly basis [16].

From the related work, machine learning techniques have been widely used for rainfall prediction. In particular, Support Vector Machine (SVM), Naïve Bayes (NB) Neural Network (NN) are the most widely used by the related work [7], [11], [12], [15]. This demonstrates the usefulness of these techniques. Hence, this study will use such techniques with two additional prediction techniques including Decision Tree and Random Forest in order to investigate the performance of these methods for rainfall prediction.

II. Materials and Methods

The data source is obtained from the Malaysia Meteorological Department and Malaysia Drainage and Irrigation Department spanning from Jan 2010 until April 2014. The location and description of the data obtained is shown in Table 1.

TABLE I
DATA SOURCE LOCATIONS

Daily Data	Station number	Station name
24 Hour Mean Temperature	48650	KLIA Sepang
24 Hour Mean Relative Humidity	48650	KLIA Sepang
24 Hour Mean Flow	2917401	SG.Langat,Kajang
Daily Totals Rainfall	2917112	Kajang , Hulu Langat
Daily Means Water Level	2917401	Sg.Langat,Kajang

The features in the dataset consist of temperature, relative humidity, flow, rainfall, and water level (Table 2).

TABLE II
FEATURES IN THE DATASET

Feature	Valid Records	Missing values
Temperatur	1581	0
Relative Humidity	1572	9
Flow	1464	117
Rainfall	1569	12
Water Level	1464	117

Table 3 shows the details of the attributes for each feature.

TABLE III
DETAILS OF THE ATTRIBUTES IN THE DATASET

Attribute name	Attribute type	Attribute measurement
Temperature	Continuous	° C
humidity	Continuous	percentage of relative humidity,%
Rainfall	Continuous	mm
River Flow	Continuous	m3/s
Water level	Continuous	ms
Class	Nominal	Rainfall-yes / rain off-no

The data pre-processing phase aims to prepare the data prior to further analysis. The weather data includes irrelevant data, noise, and incomplete instances. Pre-processing such data plays an essential role in terms of improving the performance of prediction process [17]. Hence, two tasks were performed for this purpose; cleaning and normalization. The cleaning phase will process data missing values that are represented by characters '?', '*' or negative values. Missing values has the ability to cause incorrect matches in the process of prediction [18]. Table 4 shows a sample of data with missing values.

TABLE IV
MISSING VALUES IN THE DATASET

Temperature	Humidity	Rainfall	Flow	WL
27.9	85.3	**-76.9**	3.94	22.37
27.3	86.2	**-284**	3.82	22.36
27.8	83.6	*	3.67	22.34
27.7	**-1.1**	0	10.68	22.54
28.6	73.4	0	**?**	**?**
29.3	68.3	0	**?**	**?**
29.1	67.8	5.7	**?**	**?**
28.8	67.9	11.3	**?**	**?**

In order to overcome the missing data, this study used the mean average mechanism for filling up such instances. Such mechanism aims to sum all the instances in the selected attribute then dividing the summation by the number of samples.

Normalization aims to limit the values within a specific interval. Such interval will facilitate the process of prediction where the values will be mapped onto a particular range. Normalization is essential for specific algorithms such as Neural Network and Support Vector machine [19]. In this study, the interval is set to a range between -1 to 1 based on the following formula [10].

$$y = \frac{(y_{max} - y_{min}) \times (x - x_{min})}{(x_{max} - x_{min})} + y_{min} \qquad (1)$$

Where x is the data that has to be normalized, x_{max} is the maximum value of all the input data, x_{min} is the minimum value of all the input data, y is the normalized data, y_{max} is the desired maximum value, and y_{min} is the desired minimum value.

Table 5 and 6 show the values before and after the normalization using formula (1).

TABLE V
VALUES BEFORE NORMALIZATION

Temperature	Humidity	Rainfall	Flow	WL
22.3	87.6	2.31	2.78	2.79
26.4	88.9	5.74	4.29	5.74
22.9	84.7	1.68	6.78	1.25
27.8	85.2	5.03	5.46	4.56
24.1	88.3	5.03	4.29	4.56
26.5	86.4	5.03	4.29	4.56
26.9	86.4	2.69	1.64	6.47
29.3	84.2	10.4	2.14	8.46
21.2	86.4	5.03	4.65	4.56

TABLE VI
VALUES AFTER NORMALIZATION

Temperature	Humidity	Rainfall	Flow	WL
-0.728	0.446	-0.855	-0.556	0.572
0.283	1	-0.068	0.031	0.245
-0.580	-0.756	-1	1	-1
0.629	-0.512	-0.392	1	-1
-0.283	1	-0.392	0.760	-1
0.308	1	-0.392	0.760	-1
0.407	1	-1	-1	-0.020
1	-1	1	-1	1
0	1	1	1	1

As shown in Table 6, the data has been normalized which makes it ready for further processing.

After the data has been prepared, the rainfall prediction will be performed with five techniques (Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), Neural Network (NN), and Random Forest (RF)). The evaluation of these techniques was performed using 10-fold cross validation and percentage split.

Information Retrieval metrics such as recall, precision and f-measure have been used in this study to evaluate the proposed method. The aim of Precision is to evaluate the True Positive (TP) entities that are the correctly classified entities with respect to the False Positive (FP) that are the incorrectly classified entities. It can be calculated as follows:

$$Precision = \frac{|TP|}{|TP| + |FP|} \qquad (2)$$

The aim of recall is to evaluate the True Positive with respect to the False Negative, which are the entities that not classified at all. It can be calculated as follows:

$$Recall = \frac{|TP|}{|TP|+|FN|} \quad (3)$$

However, using two values, we often cannot determine if one algorithm is superior to another. For example, if one algorithm has higher precision but lower recall than another algorithm, how can you tell which algorithm is better. A solution to this matter is by using F-measure that is the average of precision and recall calculated as follows:

$$F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

III. Result and Discussion

The experiments have been performed using Weka 3.7 that is a suite of Machine Learning software that includes various techniques. However, some of the used techniques were already installed in the software such as Naïve Bayes and Decision Tree, whereas the other techniques have been installed using plugins in the software using package manager. Note that, the experiments have been performed using two approaches; first, 10-folds cross-validation and splitting mechanism. The techniques will be discussed in terms of performance based on F-measure as follows.

The best results for Decision Tree were achieved when the training was split at 30% training and 70% testing by obtaining a score of 73.7% for F-measure (Table 7). Therefore, the Decision Tree model at 30–70 split is considered as the best model for this technique.

TABLE VII
Decision Tree Results

Model	Precision	Recall	F-measure
10-90	0.71	0.739	0.692
20-80	0.713	0.739	0.716
30-70	0.749	0.729	**0.737**
40-60	0.7	0.735	0.699
50-50	0.702	0.735	0.699
60-40	0.720	0.744	0.712
70-30	0.704	0.736	0.698
80-20	0.697	0.722	0.696
90-10	0.71	0.739	0.692

The best result for NB was achieved when the training was set at 20% training and 80% testing by obtaining 67.3% of F-measure (Table 8).

TABLE VIII
Naïve Bayes Results

Model	Precision	Recall	F-measure
10-90	0.67	0.714	0.671
20-80	0.668	0.711	**0.673**
30-70	0.654	0.71	0.658
40-60	0.673	0.725	0.67
50-50	0.690	0.731	0.671
60-40	0.677	0.720	0.655
70-30	0.684	0.728	0.652
80-20	0.687	0.716	0.644
90-10	0.724	0.728	0.648

The best result for SVM was achieved when the training was set at 20% and testing was at 80% by obtaining 67.1% of F-measure (Table 9). Therefore, 20% training – 80% testing will be considered as the best model for SVM

TABLE IX
Support Vector Machine Results

Model	Precision	Recall	F-measure
10-90	0.679	0.727	0.665
20-80	0.672	0.719	**0.671**
30-70	0.525	0.724	0.609
40-60	0.656	0.718	0.602
50-50	0.658	0.720	0.627
60-40	0.621	0.723	0.619
70-30	0.658	0.715	0.618
80-20	0.673	0.731	0.639
90-10	0.526	0.709	0.604

The best result for NN was achieved when the training was set at 60% and testing was at 40% by obtaining 74.1% of F-measure (Table 10). Therefore, 60% training – 40% testing will be considered as the best model for Neural Network.

TABLE X
Neural Network Results

Model	Precision	Recall	F-measure
10-90	0.695	0.713	0.702
20-80	0.698	0.727	0.702
30-70	0.713	0.738	0.716
40-60	0.714	0.74	0.71
50-50	0.707	0.733	0.655
60-40	0.737	0.747	**0.741**
70-30	0.698	0.730	0.680
80-20	0.731	0.741	0.735
90-10	0.685	0.722	0.690

The best result for RF was achieved when the training was at 30% and testing was at 70% by obtaining 71.9% of F-measure (Table 11). The model from 30% training – 70% testing will be considered as the best model for RF.

TABLE XI
Random Forest Results

Model	Precision	Recall	F-measure
10-90	0.704	0.733	0.704
20-80	0.711	0.729	0.717
30-70	0.715	0.738	**0.719**
40-60	0.699	0.72	0.706
50-50	0.711	0.733	0.716
60-40	0.694	0.720	0.699
70-30	0.710	0.732	0.715
80-20	0.645	0.681	0.653
90-10	0.689	0.715	0.693

For the cross-validation approach, NN has outperformed the other techniques by obtaining the highest scores for Precision (72.1%), F-measure (72.5%) and Recall (74.4%). RF outperformed NB and SVM by achieving 70.7% of F-measure and Precision (70.1%). Finally, the lowest value of F-measure has been obtained by SVM (Fig. 1).

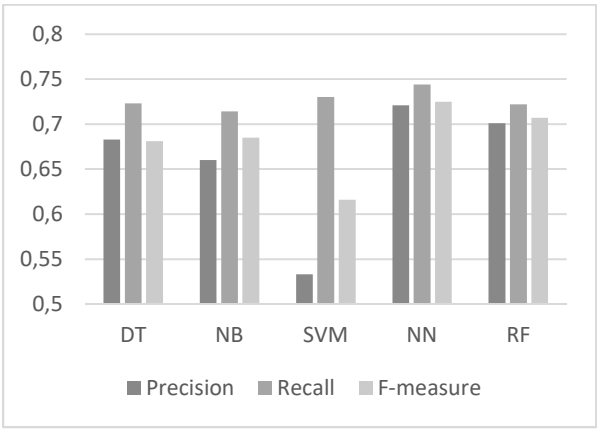

Fig. 1 Comparison of cross-validation approach among the five techniques

On the other hand, in terms of percentage split (Fig. 2), the effect of splitting approach on techniques performance is varying from technique to the other according to the percentage split between training and testing. The differences were affected by the behavior of each technique.

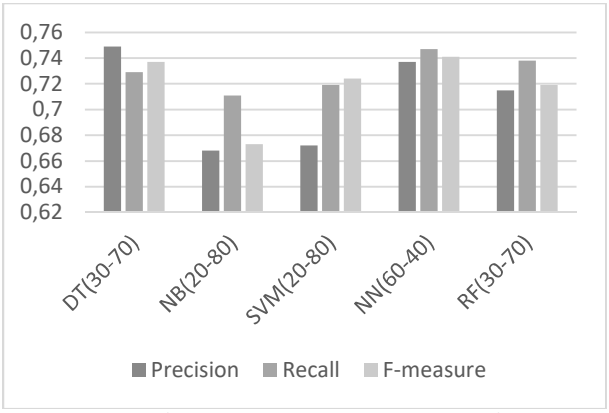

Fig. 2 Comparison of percentage split approach among the five techniques

The higher F-measure is 74.1% from NN using 60% training data and 40% testing data. This indicates that NN depends on more data for training to ensure a good model. This is the behavior of NN that needs to adjust the optimal weights for the training data. The portion of training data in this situation may be not enough to get an optimal weight to fit the data. Decision Tree comes in the second with F-measure 73.7% using 30% training data and 70% testing data. Decision Tree builds the tree based on the rules that represent the training data. Decision Tree looks for the feature that has more ability (more information) to split the data in order to build the tree.

Random Forest behaves similarly to Decision Tree; however Random Forest builds an ensemble of trees (i.e. forest) and the main principle behind ensemble methods is that a group of "weak learners" can come together to form a "strong learner". Random Forest is a combination of separate trees. Thus, each tree is a weak learner. However, when the trees are ensembled in a Random Forest, the end model is a strong learner. With the ensemble strength, Random Forest achieved 72.3% F-measure and the model is based on 30% training data and tested on 70% of testing data.

Naïve Bayesian classifiers behavior assumes attributes have independent distributions so it is not sensitive to irrelevant features. Naïve Bayes models also use the method of maximum likelihood, therefore it required only a small amount of training data for prediction. The best model for Naïve Bayes using 20% training data and 80% testing data scored 67.3% for F-measure. By using kernel functions, SVM is able to learn high-quality decision boundaries that can be efficiently generalized onto test data, so it can learn and find optimal hyperplane using a small set of training data. For the rainfall data, the SVM model based on 20% training data and 80% testing data achieved 67.1% F-measure score.

Comparing these five techniques performances for rainfall prediction, Decision Tree, and Random Forest are the top performers. This is supported by the fact that although these models were trained on a low portion of training data, the models were able to predict the higher portion of testing data with the top F-measure scores. Compared with Support Vector Machine and Naive Bayes that were trained on a small portion of training data and predict the higher portion of test data but scored lower F-measures. Neural Network it is an efficient method but it needs a large portion of training data to train in order to predict very small portion of test data.

All techniques produced low predictions' scores between 63% and 75 %. There are three possible reasons for this low performance; first are the sizes of the datasets in these experiments. This study used data collected between January 2010 and April 2014(less than 5 years). On the other hand, previous research [7], [11], and [12] used data from 10 to 50 years for rainfall prediction. The minimum period (10 years) is more than double of the data used in this research (less than 5 years). The longer period equals to more data and produces a more informative model.

Second are the 117 missing values for water flow and water level. Third is the lack of other relevant features like wind speed which were used in previous research. Therefore, to improve the results, a larger collection dataset of at least 10 years is needed with more relevant types of data and better methods to estimate the missing values.

As a final evaluation, we further analyzed the experiments for all classifiers using 90% training data and 10% testing data. Since our data is quite small, we focused on the model that utilizes the largest amount of instances to train the model (10% training data and 90% testing data) (Table 12).

TABLE XII
CORRECTLY CLASSIFIED INSTANCES AT 10% TRAINING

Technique	**Correctly classified instances at 10% training data**
SVM	1034
NB	1015
RF	**1043**
DECISION TREE	1039
ANN	1015

From Table 12 we can conclude that with small training data (10%) from 1581 total, Random Forest correctly classified 1043 instances (highest number of correct instances), therefore Random Forest is in the forefront of the five techniques in this study.

IV. CONCLUSIONS

This research has successfully accomplished the objectives where five classification techniques (Naïve Bayes, Decision Tree, Support Vector Machine, Neural Network and Random Forest) were performed for Malaysian rainfall prediction. The main objective of this study is to identify the best technique for rainfall prediction. Hence, after applying the five techniques, a comparative analysis has been performed in order to determine the most appropriate technique. The experimental results showed that for Rainfall prediction, Decision Tree, and Random Forest perform well because of their abilities to train on little data and predict the higher portion of data with higher F-measure. Support Vector Machine and Naive Bayes also trained on a small portion of data to predict higher portion but with lower F-measure. Neural Network it is an efficient method but it needs a large portion of training data to predict the very small portion of testing data. In addition, we can conclude that with small training data (10%) from 1581 instances Random Forest correctly classified 1043 instances. This result put Random Forest in the forefront of the five techniques we have been used.

For future work, the following suggestions can be considered; Combining two or more prediction algorithms has the ability to enhance the process of predicting; Use more valuable features that can generalize or discriminate the classes has a significant impact on the effectiveness; Exploit the rainfall prediction has a significant impact on predicting flowed where there is a direct correlation between the rainfalls and flowed; Use more dataset and explore more areas and locations in the world would be a valuable idea.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Malaysia (Grant Code FRGS/2/2013/ICT02/UKM/02/).

REFERENCES

[1] A. Rajaraman and J. D. Ullman, *Mining of massive Datasets*, 20th ed. Cambridge: Cambridge University Press (Virtual Publishing), 2012.

[2] D. J. Hand, H. Mannila, P. Smyth, and D. J. H, *Principles of data mining (Adaptive computation and machine learning)*. Cambridge, MA: Bradford Books, 2001.

[3] C. L. Wu and K. W. Chau, "Prediction of rainfall time series using modular soft computing methods,"*Engineering Applications of Artificial Intelligence*, vol. 26, no. 3, pp. 997–1007, Mar. 2013.

[4] S. Rani and G. Sikka, "Recent techniques of clustering of time series data: A survey," *International Journal of Computer Applications*, vol. 52, no. 15, pp. 1–9, Aug. 2012.

[5] H. S. Badr, B. F. Zaitchik, and A. K. Dezfuli, "A tool for hierarchical climate regionalization," *Earth Science Informatics*, vol. 8, no. 4, pp. 949–958, May 2015.

[6] I. Panel, C. Change, and P. Ivonne, Climate change 2013: The physical science basis: Working group I contribution to the fifth assessment report of the intergovernmental panel on climate change, Intergovernmental Panel on Climate Change, Ed. Cambridge: Cambridge University Press, 2014.

[7] K. Abhishek, A. Kumar, R. Ranjan, and S. Kumar, "A rainfall prediction model using artificial neural network," in Control and System Graduate Research Colloquium (ICSGRC), 2012, IEEE, 2012. [Online]. Available: 10.1109/ICSGRC.2012.6287140. Accessed: Nov. 6, 2016.

[8] R. Venkata Ramana, B. Krishna, S. R. Kumar, and N. G. Pandey, "Monthly rainfall prediction using Wavelet neural network analysis," *Water Resources Management*, vol. 27, no. 10, pp. 3697–3711, Jun. 2013.

[9] B. Wang et al., "Rethinking Indian monsoon rainfall prediction in the context of recent global warming," Nature Communications, vol. 6, p. 7154, May 2015.

[10] K. K. Htike and O. O. Khalifa, "Rainfall forecasting models using focused time-delay neural networks,"*International Conference on Computer and Communication Engineering (ICCCE'10)*, May 2010.

[11]D. Gupta and U. Ghose, "A comparative study of classification algorithms for forecasting rainfall," *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Sep. 2015.

[12] J. Joseph and R. T K, "Rainfall prediction using data mining techniques," *International Journal of Computer Applications*, vol. 83, no. 8, pp. 11–15, Dec. 2013.

[13] "The Oryx resource guide to El Nino and la Nina," *Choice Reviews Online*, vol. 40, no. 04, pp. 40–2226–40–2226, Dec. 2002.

[14] J. Kajornrit, K. W. Wong, and C. C. Fung, "Rainfall prediction in the northeast region of Thailand using modular fuzzy inference system," *2012 IEEE International Conference on Fuzzy Systems*, Jun. 2012.

[15] V. B. Nikam and B. B. Meshram, "Modeling rainfall prediction using data mining method: A Bayesian approach," *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*, Sep. 2013.

[16] M. Kühnlein, T. Appelhans, B. Thies, and T. Nauss, "Improving the accuracy of rainfall rates from optical satellite sensors with machine learning — A random forests-based approach applied to MSG SEVIRI,"*Remote Sensing of Environment*, vol. 141, pp. 129–143, Feb. 2014.

[17] D. Isa, L. H. Lee, V. P. Kallimani, and R. RajKumar, "Text document Preprocessing with the Bayes formula for classification using the support vector machine," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1264–1272, Sep. 2008.

[18] R. S. V. Teegavarapu and V. Chandramouli, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records," *Journal of Hydrology*, vol. 312, no. 1-4, pp. 191–206, Oct. 2005.

[19] S. S. Monira, Z. M. Faisal, and H. Hirose, "Comparison of artificially intelligent methods in short term rainfall forecast," *2010 13th International Conference on Computer and Information Technology (ICCIT)*, Dec. 2010.