Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia

Muchamad Taufiq Anwar
Faculty of Information Technology
Universitas Stikubank
Semarang, Indonesia
taufiq@edu.unisbank.ac.id

Edy Winarno
Faculty of Information Technology
Universitas Stikubank
Semarang, Indonesia
edywin@edu.unisbank.ac.id

Wiwien Hadikurniawati
Faculty of Information Technology
Universitas Stikubank
Semarang, Indonesia
wiwien@edu.unisbank.ac.id

Wahyu Widiyatmoko
Geography Education
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
wahyu.widiyatmoko@ums.ac.id

2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-7281-8406-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISRITI51436.2020.9315460

Abstract— **Rain prediction is a crucial topic that continues to gain interest across the globe. Rain has a massive impact on various aspects of human life such as in agriculture, health, transportation, etc, and also some natural disasters. Various impacts of rain on human life prompts us to build a model to understand and predict rain to provide early warning for various use cases in various fields. Previous research on rain modeling using Data Mining (DM) techniques had suffered from low accuracy caused by the limited availability of the training data and their meteorological attributes. This research aims to address those issues by building the rain model using a richer and more abundant rain data in Indonesia. Four DM techniques are used and compared in this research i.e. the C4.5/J48, Random Forest (RF), Naïve Bayes (NB), and Multilayer Perceptron (MLP). The experimental results showed that the MLP and J48 algorithm can provide the best accuracy (up to 78,4%), which is better than previous research. Other key findings in this research include: (a) the selection of DM techniques has little effect on the model accuracy; (b) a larger training dataset generally improves model accuracy and a larger test dataset is necessary to get a representative real-world test accuracy, and (c) the two most influential attributes in rain modeling are the relative humidity and the minimum temperature, and we suggest to include cloud condensation nuclei in the next research to complete the model.**

Keywords— Rain prediction; Data Mining; classification; J48, Random Forest; Naïve Bayes, Multilayer Perceptron

I. INTRODUCTION

Rain prediction is a crucial topic that continues to gain interest across the globe. Rain has a massive impact on various aspects of human life such as agriculture, health, transportation, etc. Other than that, rain also affects natural disasters such as landslides and floods. The various impacts of rain on human life prompts us to create a model to understand and predict rain to provide early warning in various fields/needs such as agriculture, transportation, etc. Rain modeling can be done by applying Data Mining (DM) / Machine Learning (ML) algorithms to historical weather data that has been captured by meteorological stations that are scattered in various locations. Literature had shown that DM / ML can be applied to weather prediction and forecasting [1][2]. Previous research on rain modeling using Data Mining technique conducted in Lahore City, Pakistan [3] has low accuracy of only 40% on rain prediction which is caused by missing values, a limited set of attributes, and low rain occurrence in the study area which results in a low number of datasets in the 'rain' class. Other research conducted in Malaysia uses daily weather data in 52 months and has a maximum accuracy of 74,1% and had pointed out the need to add more meteorological variables and also more datasets to increase the accuracy [4]. It is also generally recommended to explore new location as the study area [3], [4] as the location where the data is taken can affect the model accuracy [5]. A recent literature review on rain prediction using DM techniques showed that ongoing research should focus on improving model accuracy [5]. Based on the aforementioned situation, this research aims to address the weaknesses from previous research to increase the model performance by exploring more meteorological variables, more dataset in 'rain' class, and by using larger datasets in general. Indonesia is suited for this research since there are a high number of rainy days in Indonesia which will contribute to more abundant rain datasets.

Previous research uses some meteorological attributes such as temperature [3], [4], humidity [3], [4], and wind speed [3]. DM methods used in previous research are Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Artificial Neural Network ANN / Multilayer Perceptron (MLP) [3], [4], and K-Nearest Neighbor (KNN) [3]. This research includes some additional attributes as provided by the Indonesian Meteorology, Climatology, and Geophysics Agency (BMKG, bmkg.go.id) i.e the minimum temperature, maximum temperature, average temperature, average relative humidity, sun exposure time, maximum wind speed, and average wind speed. This study will use and compare the performance of four top-performing Data Mining algorithms known to date, i.e. the J48, Random Forest, Naïve Bayes, and Multilayer Perceptron. It is worth mentioning that other than the Data Mining techniques, other approaches might be used for rain prediction, such as Exponential Smoothing [6].

A. Data Mining Techniques

Data Mining / Machine Learning is automated learning to find patterns in data. Data Mining / Machine Learning approach had been used in rain prediction using methods such as J48 [7] and Artificial Neural Networks [8]. Some research may use the weather prediction to be linked to a certain phenomenon such as Dengue Fever [9] and agriculture/food [10][11]. In this research, Random Forest, Naïve Bayes, and Multilayer Perceptron will be used. Also, rain modeling using the J48 from previous research [7] will be used as a comparison. J48 is a well-performing decision tree modeling algorithm and had been used in many areas

such as in wildfire modeling [12]. In principle, J48 creates a tree based on the value of entropy and information gain for each attribute. The formula for entropy and information gain is shown in (1) and (2).

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (1)$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

B. Random Forest

Research [13] had shown that a decision tree is very suitable for rain prediction. But, one disadvantage of a single decision tree, such as C4.5/J48, is overfitting. Random Forest (RF) handles this shortcoming by creating multiple numbers of trees, by randomly selecting the training data (bootstrap sampling), and also by randomly selecting the attributes to create branching in the decision trees. To make the prediction, a set of attributes will be fed into the trees to get the prediction from each tree, and the final prediction will be based on the voting from the predicted value from those trees. This use of multiple trees is called ensemble learning. The Random Forests algorithm was devised by Breiman in 2001 [14]. Since then, RF had been popularly used in many areas including economics [15]. The branching of the trees is based on Gini impurity. Gini impurity is the probability measurement of a random sample to be incorrectly classified when a new instance was randomly classified based on the distribution of class labels in the data set. The calculation of Gini impurity for a set of items with J classes, supposing $i \in \{1,2,\ldots,J\}$, and f_i be the fraction of items labeled with class i in the set, the formula is shown in (3) [15].

$$I_G(f) = \sum_{i=1}^{J} f_i(1 - f_i) = \sum_{i \neq k} f_i f_k \quad (3)$$

C. Naïve Bayes

Naïve Bayes (NB) is a probabilistic modeling prediction/classification based on the Bayesian theorem with the 'naïve' assumption of independence among predictors. The conditional probability of an occurrence of event A when event B occurs is determined by (4) [16]. The final predicted class would then determined by the class with the highest probability using the argmax function. NB had been used in many areas including breast cancer prediction [16] and fake news detection [17] with good accuracy, although other methods such as Artificial Neural Network [16] and K-Nearest Neighbour [18] might perform better.

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)} \quad (4)$$

where

P(A|B) = the probability of the occurrence of event A when event B occurs

P(A) = the probability of the occurrence of A

P(B|A) = the probability of the occurrence of event B when event A occurs

P(B) = the probability of the occurrence of B.

D. Multilayer Perceptron

A multilayer perceptron (MLP) is a type of Artificial Neural Network (ANN) which consists of several layers of neurons where the learning is accomplished by forward-feeding, backpropagation, and an adaptive learning rate. The MLP structure usually consists of at least three layers [19], one input layer where ANN receptors receive external data, one output layer where the solution to the problem is obtained (in this case whether the class is 'rain' or 'no rain'). In between those two layers, there is at least one intermediate layer called the hidden layer. An example of the structure of an ANN/MLP diagram is shown in Fig. 1. The ANN itself is a popular Machine Learning technique with numerous types and usages, such as the Convolutional Neural Network (CNN) used for face-recognition in an attendance system [20].

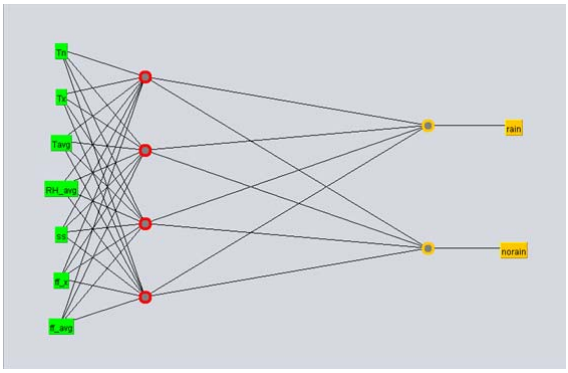

Fig. 1. An example of an MLP structure (as used in this research)

II. Materials and Methods

The method of this research is shown in Fig. 2. Daily historical weather data were obtained from the BMKG website for the Tanjung Mas meteorological station, in Semarang City, Indonesia, spanning from 2013 to 2019 with a total of 2526 rows of data. The original data taken from BMKG consisted of 11 attributes i.e: Date, Tn: Minimum temperature (°C), Tx: Maximum temperature (°C), Tavg: Average temperature (°C), RH_avg: Average humidity (%), RR: Rainfall (mm), ss: The duration of the sun (hours), ff_x: Maximum wind speed (m/s), ddd_x: Wind direction at maximum speed (°), ff_avg: Average wind speed (m/s), and ddd_car: Most wind directions (°). For this research, only 8 attributes from those 11 attributes were used, i.e. minimum temperature, maximum temperature, average temperature, average relative humidity, sun exposure time, maximum wind speed, and average wind speed. The wind direction attribute (ddd_x) was not used because the numerical scale is problematic to use, for example, 0° and 359° seems separated far apart yet in reality it is very close in terms of wind direction. The date and most wind direction (ddd_car) attributes which have nominal value is also removed. Next, class labeling is done by evaluating the value in the RR (rainfall) attribute, if RR> 0 then class = 'rain'; otherwise, class = 'no rain'. The RR attribute is then intentionally removed for these classification tasks because it has been replaced by the target class. Data cleaning is performed to remove entries with one or more missing values. The final dataset has all numerical attributes, except for the nominal target class. The final attributes in the dataset are shown in Table I.

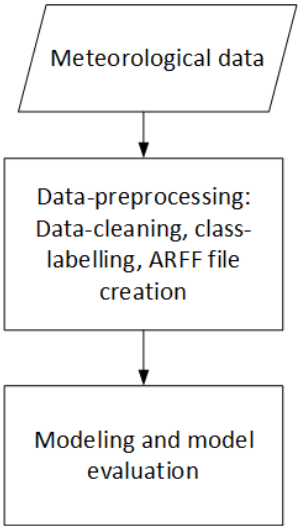

Fig. 2. Research methods

TABLE I. ATTRIBUTES OF THE METEOROLOGICAL DATA

Attribute	Data type	Description
Tn	Numeric	Minimum temperature
Tx	Numeric	Maximum temperature
Tavg	Numeric	Average temperature
RH_avg	Numeric	Average Humidity (%)
ss	Numeric	Sun exposure time (hours)
ff_x	Numeric	Maximum wind speed (m/s)
ff_avg	Numeric	Average wind speed (m/s)
rainy	Nominal	Target class

After cleaning 449 (18%) entries with missing values, 2077 rows of data are used as the final dataset. This dataset consists of 877 rows of data (42%) for the 'rain' class and 1200 data (58%) for the 'no rain' class. The data is then stored in CSV format and then converted to an ARFF file format to be processed using the Waikato Environment for Knowledge Analysis (WEKA) software [21]. Experiments were carried out using the J48, Naïve Bayes, Random Forest, and Multilayer Perceptron functions in the WEKA classification tab. The normalization of the numerical attributes is automatically performed when building the MLP model. Evaluation of the performance of the training model is done using the split method, full training data, and 10-fold cross-validation. Model performance evaluation is conducted by using several measurements ie. Accuracy, Precision, Recall, and F-measure. All of these measurements are based on False Positive (FP), False Negative (FN), True Positive (TP), and True Negative (TN). Precision and Recall is a necessary measure to show the model's performance for a particular class (which is especially useful in a dataset with imbalanced class), which can not be told by Accuracy. Accuracy is defined as the total of correctly classified instances (TP and TN) divided by all test instances. The formula for Accuracy is shown in (5).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

Precision shows the portion of positive correctly classified instances (TP) against total instances which are predicted positive (TP and FP). The formula for Precision is shown in (6).

$$Precision = \frac{TP}{(TP + FP)} \quad (6)$$

Recall shows the portion of positive correctly classified instances (TP) against total instances which are positive in reality (TP and FN). The formula for Recall is shown in (7).

$$Recall = \frac{TP}{(TP + FN)} \quad (7)$$

F-measure can be seen simply as the average from Precision and Recall which tells the overall model performance. The formula for F-measure is shown in (8).

$$F - measure = \frac{Precision * Recall * 2}{(Precision + Recall)} \quad (8)$$

III. RESULTS AND DISCUSSIONS

A. Performance Comparison of DM Techniques

Table II showed the model performance using 10-fold cross-validation with entries with missing values are omitted, whereas Table III shows the model performance with entries with missing values included. The result showed that MLP has the best performance followed by J48, although the difference is thin. When we experiment to include entries with missing values, we found that compared to other methods, NB can slightly benefit from non-complete entries. Overall, removing entries with missing values gives only a very slight improvement in model performance. As all the 4 methods give a nearly identical performance, we argue that the model performance is more influenced by the data rather than the method used. Table II and Table III also showed that the Precision and Recall are nearly identical which means that our algorithms had classified nearly the same number of instances as FP and FN. Table II and Table III showed that all of the measurement in the 'rain' class is slightly lower than those of the 'norain' class which means that there is slightly worse prediction ability in the 'rain' class. This might be caused by a relatively lower number of datasets in the 'rain' class (41%). In terms of accuracy, the model created by MLP has the best accuracy, followed by J48, as shown in Table IV.

TABLE II. MODEL PERFORMANCE USING ALL 7 ATTRIBUTES

Method	Class	Precision	Recall	F-measure
J48	rain	0.733	0.726	0.730
	norain	0.801	**0.807**	0.804
RF	rain	0.633	0.633	0.633
	norain	0.732	0.732	0.732
NB	rain	0.708	0.708	0.708
	norain	0.787	0.787	0.787
MLP	rain	0.731	0.754	0.742
	norain	**0.816**	0.798	**0.807**

TABLE III. MODEL PERFORMANCE USING ALL 7 ATTRIBUTES (ENTRIES WITH MISSING VALUES ARE INCLUDED)

Method	Class	Precision	Recall	F-measure
J48	rain	0.730	0.746	0.738
	norain	**0.801**	0.788	0.794
RF	rain	0.637	0.654	0.646
	norain	0.728	0.713	0.720
NB	rain	0.724	0.728	0.726
	norain	0.789	0.786	0.788
MLP	rain	0.741	0.708	0.724
	norain	0.783	**0.809**	**0.796**

TABLE IV. MODEL ACCURACY USING 10-FOLD CROSS-VALIDATION

Method	Model accuracy
J48	77.3
RF	76.8
NB	75.4
MLP	**77.9**

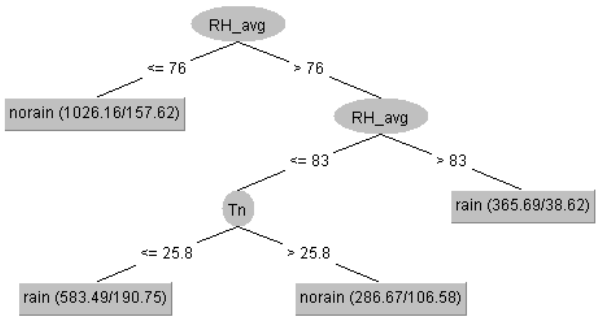

Fig. 3. Simplified decision tree produced by J48

Table IV shows that J48 has better accuracy than RF, thus we argue the "low" accuracy of the model is not caused by model overfitting. As all of the experiments with various methods show roughly similar accuracy of the training model, which is around 75.4-77.9%, we argue that the methods used for modeling have little effect on model accuracy but instead, the model accuracy is more affected by the training data.

From the previous research on rain model using J48 [7], we are interested to see the effect of the minimum objects per leaf to model accuracy using the J48 method. The decision tree produced by the J48 algorithm from previous research [7] is shown in Fig. 3. The result in Table V showed that as the number of minimum objects is increased, J48 gains slight accuracy improvement which is the result of more pruning of the tree and thus makes the model suffer less from overfitting, as J48 is known to have this weakness on overfitting [22]. Overall, the result of this study has achieved a higher accuracy and F measure than previous research [3], [23]. The relatively higher accuracy achieved by the J48 method is in line with other studies which stated that the Decision Tree model is better than the other prediction models [24]. Although, in this research, we find that MLP can have a slightly better performance.

TABLE V. J48 MODEL'S ACCURACY USING A DIFFERENT NUMBER OF THE MINIMUM OBJECT PER LEAF

Min object per leaf	Accuracy
2	76.9
5	76.9
10	77.4
30	77.7
50	78.1

B. The Models' Accuracy Against Testing Data

In addition to intrinsic testing in the training model, the model is also tested against real weather data in 2020. Testing data consists of weather data from January to April 2020 consisting of 121 rows of data with 68 (56%) 'rain' data and 39 (44%) 'no rain' data. Table VI showed that the accuracy against the testing set is slightly lower than the accuracy in training data. This means that our testing data are quite representative of the training data despite it only consist of four months of data. Monthly data from January to April have similar class proportions but provide different accuracy results, especially between January and March which have striking differences. This might be caused by the varied values of the attributes and not because of the uneven class distribution since the class proportion is the same for those two months. When tested against actual test data in 2020, J48 performs best while RF performs worst. The difference in accuracy among months might be caused by the low number of datasets (30 entries). Thus we suggest that future research should use a larger test dataset to have a more consistent test accuracy. It is also suggested that we should not rely only on one method when building a model. Also, the model created by using different datasets would show different performance. If we want to have a 'universal' weather model, we suggest using broader data covering larger temporal and spatial scale to have a richer data and hoped that this will create a more generalized pattern/model. But if we want to have a good model used for a certain location, we should focus on having more datasets in that particular location.

TABLE VI. MODEL'S ACCURACY AGAINST TESTING DATA

	Test accuracy (%)				
Method	Jan (58% rain)	Feb (62% rain)	Mar (58% rain)	Apr (56% rain)	Jan-Apr (56% rain)
J48	85.7	76.0	**67.9**	**77.8**	**76.6**
RF	78.6	56.0	64.3	66.7	66.4
NB	**96.4**	76.0	57.1	63.0	72.9
MLP	78.6	**80.0**	60.7	74.1	72.9

C. The Effect of the Amount of Training Data to Models' Accuracy

Table VII shows the model accuracy using different percentage split of training data. In general, the more data training is used, the better the model accuracy is, although

there might be minor fluctuation. Overall, MLP has the best accuracy among other methods in all test cases, whether the number of training data is low or high, except when using the full training dataset. RF can have 100% accuracy when using the full training dataset, but this may not necessarily reflect its real performance on real test data. On the other hand, RF performs worst when the number of data is low. But when compared to the result from 10-fold cross-validation that uses 90% of training data, the accuracy may fall slightly. As stated in [13] an increase in the size of the training set, accuracy first increases but then decreases after a limit. We argue that this decline in the accuracy might be caused by the variation/anomalies within the training data, but this should be investigated by further research.

TABLE VII. ACCURACY OF THE TRAINING MODEL RELATED TO THE PERCENTAGE SPLIT OF THE TRAINING DATA

	Model Accuracy (%)				
Method	**10% split**	**15% split**	**50% split**	**85 split**	**100% training**
J48	75.1	76.0	78.3	79.2	83.0
RF	69.1	68.8	69.9	79.5	**100**
NB	74.8	75.0	75.1	76.6	75.7
MLP	**76.2**	**77.7**	**78.4**	**79.8**	78.9

D. The Contribution od Attributes to Models' Accuracy

As found in the previous study [7], the two most influential attributes to rain modeling are the relative humidity and the minimum temperature. Based on that finding, we experiment to create the model by using only those two attributes. The result in Table VIII showed that by only using the two attributes, the model gained a slightly higher accuracy than that with the full 7 attributes, except for the RF. This result challenged the idea that adding more attributes will lead to better model accuracy, as suggested by previous research [3], [4]. The slight decline in the RF model might be caused by the fact that it can not build more diverse trees by using only two attributes. The result in Table IX also showed that performances on the 'rain' classes remain slightly lower than those of 'norain' class which means that the model for 'rain' is more in the need of additional affecting attributes. Maybe the attributes we're looking for to complete the rain model is not yet available in previous nor this research. It is known that the process of the rain to happen is by a process of condensation which is related to humidity and minimum temperature (as stated in the finding of the previous study [7]) but also can be affected by the presence of a nucleus that "facilitates" condensation. These nuclei are known as the cloud condensation nuclei which plays an important role in building accurate climate modeling [25]. These nuclei can be in the form of dust, smoke, salt, etc. These nuclei can reduce the need for lower RH numbers for condensation to occur. As stated in [26] the concentrations of Cloud Condensation Nuclei are highly influential to the intensity of the precipitation. Thus, the inclusion of nuclei into the model is hoped to complete the model and thus increase its accuracy. We recommend that future research should include this attribute in the rain model.

TABLE VIII. MODELS' ACCURACY BY USING ONLY TWO ATTRIBUTES

	Model accuracy (%)	
Method	**Full attributes**	**Two attributes only**
J48	77.3	**78.4**
RF	**76.8**	76.6
NB	75.4	**76.2**
MLP	77.9	**78.1**

TABLE IX. MODELS' PERFORMANCE BY USING ONLY TWO ATTRIBUTES

Method	**Class**	**Precision**	**Recall**	**F-measure**
J48	rain	0.756	0.723	0.739
	norain	0.804	**0.829**	**0.816**
RF	rain	0.723	0.722	0.723
	norain	0.797	0.798	0.798
NB	rain	0.691	0.791	0.738
	norain	**0.829**	0.741	0.783
MLP	rain	0.743	0.737	0.740
	norain	0.809	0.814	0.811

IV. CONCLUSION

Rain prediction models are very useful for human life. This study compares four DM techniques used for the rain prediction model, i.e. J48, RF, NB, and MLP. Results showed that MLP and J48 algorithm can provide the best accuracy (up to 78,4%) compared to other algorithms, although the difference is small. We had achieved better accuracy than previous research. Other key findings in this research include: (a) the selection of DM techniques has little effect on the model accuracy; (b) a larger training dataset generally improves model accuracy and a larger test dataset is necessary to get a representative real-world test accuracy, and (c) the two most influential attributes in rain modeling are the relative humidity and the minimum temperature. Future research that tries to improve the model accuracy should add cloud condensation nuclei to complete the model and pay attention to the possibilities of anomalies in the training dataset. Future research may also compare models from different locations and do parameter tuning of the model creation to increase its performance.

REFERENCES

[1] M. R. Mahmood, R. K. Patra, R. Raja, and G. R. Sinha, "A novel approach for weather prediction using forecasting analysis and data mining techniques," in *Innovations in Electronics and Communication Engineering*, Springer, 2019, pp. 479–489.

[2] C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, and H. S. Kim, "Development of heavy rain damage prediction model using machine learning based on big data," *Adv. Meteorol.*, vol. 2018, 2018.

[3] S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz, "Rainfall Prediction in Lahore City using Data Mining Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 4, pp. 254–260, 2018.

[4] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for Malaysian rainfall prediction," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 1148–1153, 2016.

[5] S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz, "Rainfall Prediction using Data Mining Techniques: A Systematic Literature Review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 5, pp. 143–150, 2018.

[6] K. D. Hartomo, S. Y. J. Prasetyo, M. T. Anwar, and H. D. Purnomo, "Rainfall Prediction Model Using Exponential Smoothing Seasonal Planting Index (ESSPI) For Determination of Crop Planting Pattern," in *Computational Intelligence in the Internet of Things*, IGI Global, 2019, pp. 234–255.

[7] M. T. Anwar, S. Nugrohadi, V. Tantriyati, and V. A. Windarni, "Rain Prediction Using Rule-Based Machine Learning Approach," *Adv. Sustain. Sci. Eng. Technol.*, vol. 2, no. 1, 2020.

[8] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, "A comprehensive survey of data mining techniques on time series data for rainfall prediction," *J. ICT Res. Appl.*, vol. 11, no. 2, pp. 168–184, 2017.

[9] N. Agarwal, S. R. Koti, S. Saran, and A. S. Kumar, "Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India," *Curr. Sci.*, vol. 114, no. 11, pp. 2281–2291, 2018.

[10] P. S. Tayde, B. K. Patil, and R. A. Auti, "Applying Data Mining Technique to Predict Annual Yield of Major Crops," *Int. J.*, vol. 2, no. 2, 2017.

[11] U. K. Dey, A. H. Masud, and M. N. Uddin, "Rice yield prediction model using data mining," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2017, pp. 321–326.

[12] M. T. Anwar, H. D. Pumomo, S. Y. J. Prasetyo, and K. D. Hartomo, "Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 409–415.

[13] R. S. Kumar and C. Ramesh, "A study on prediction of rainfall using datamining technique," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, vol. 3, pp. 1–9.

[14] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[15] W. Long, L. Song, and L. Cui, "Relationship between capital operation and market value management of listed companies based on random forest algorithm," *Procedia Comput. Sci.*, vol. 108, pp. 1271–1280, 2017.

[16] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019.

[17] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017, pp. 900–903.

[18] E. Zuliarso, M. T. Anwar, K. Hadiono, and I. Chasanah, "Detecting Hoaxes in Indonesian News Using TF/TDM and K Nearest Neighbor."

[19] B. T. Pham, D. T. Bui, I. Prakash, and M. B. Dholakia, "Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS," *Catena*, vol. 149, pp. 52–63, 2017.

[20] E. Winarno, I. H. Al Amin, H. Februariyanti, P. W. Adi, W. Hadikurniawati, and M. T. Anwar, "Attendance System Based on Face Recognition System Using CNN-PCA Method and Real-time Camera," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2019, pp. 301–304.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[22] M. Z. F. Nasution, O. S. Sitompul, and M. Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4. 5 classification," in *J. Phys. Conf*, 2018, vol. 978, no. 012058, pp. 10–1088.

[23] N. Z. M. Safar, A. A. Ramli, H. Mahdin, D. Ndzi, and K. M. N. K. Khalif, "Rain prediction using fuzzy rule based system in North-West Malaysia," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1572–1581, 2019.

[24] N. W. Zamani and S. S. M. Khairi, "A comparative study on data mining techniques for rainfall prediction in Subang," in *AIP Conference Proceedings*, 2018, vol. 2013, no. 1, p. 20042.

[25] P. Liu *et al.*, "Resolving the mechanisms of hygroscopic growth and cloud condensation nuclei activity for organic particulate matter," *Nat. Commun.*, vol. 9, no. 1, pp. 1–10, 2018.

[26] P. J. Marinescu, S. C. van den Heever, S. M. Saleeby, S. M. Kreidenweis, and P. J. DeMott, "The Relative Roles of Lower- and Middle-Tropospheric Cloud Condensation Nuclei on Mature MCS Precipitation Rates," *AGUFM*, vol. 2016, pp. A52C--06, 2016.