Machine Learning Techniques for Short-Term Rain Forecasting System in the Northeastern Part of Thailand

Lily Ingsrisawang, Supawadee Ingsriswang, Saisuda Somchit, Prasert Aungsuratana, and Warawut Khantiyanan

***Abstract*—**This paper presents the methodology from machine learning approaches for short-term rain forecasting system. Decision Tree, Artificial Neural Network (ANN), and Support Vector Machine (SVM) were applied to develop classification and prediction models for rainfall forecasts. The goals of this presentation are to demonstrate (1) how feature selection can be used to identify the relationships between rainfall occurrences and other weather conditions and (2) what models can be developed and deployed for predicting the accurate rainfall estimates to support the decisions to launch the cloud seeding operations in the northeastern part of Thailand. Datasets collected during 2004-2006 from the Chalermprakiat Royal Rain Making Research Center at Hua Hin, Prachuap Khiri khan, the Chalermprakiat Royal Rain Making Research Center at Pimai, Nakhon Ratchasima and Thai Meteorological Department (TMD). A total of 179 records with 57 features was merged and matched by unique date. There are three main parts in this work. Firstly, a decision tree induction algorithm (C4.5) was used to classify the rain status into either rain or no-rain. The overall accuracy of classification tree achieves 94.41% with the five-fold cross validation. The C4.5 algorithm was also used to classify the rain amount into three classes as no-rain (0-0.1 mm.), few-rain (0.1- 10 mm.), and moderate-rain (>10 mm.) and the overall accuracy of classification tree achieves 62.57%. Secondly, an ANN was applied to predict the rainfall amount and the root mean square error (RMSE) were used to measure the training and testing errors of the ANN. It is found that the ANN yields a lower RMSE at 0.171 for daily rainfall estimates, when compared to next-day and next-2-day estimation. Thirdly, the ANN and SVM techniques were also used to classify the rain amount into three classes as no-rain, few-rain, and moderate-rain as above. The results achieved in 68.15% and 69.10% of overall accuracy of same-day prediction for the ANN and SVM models, respectively. The obtained results illustrated the comparison of the predictive power of different methods for rainfall estimation.

***Keywords*—**Machine learning, decision tree, artificial neural network, support vector machine, root mean square error.

L. Ingsrisawang and P. Aungsuratana are with Department of Statistics, Kasetsart University, Faculty of Science, Kasetsart University, Bangkok, Thailand (e-mail: fscilli@ku.ac.th).

Supawadee Ingsriswang is with National Center for Genetic Engineering and Biotechnology, Bangkok, Thailand.

P. Aungsuratana and W. Khantiyanan are with Bureau of the Royal Rainmaking and Agriculture Aviation, Bangkok, Thailand.

I. Introduction

THE northeastern part of Thailand is an arid region with varied rainfall. To enhance the precipitation in this area, a number of cloud seeding operations have been conducted by the Royal Rain Making Project. Since there is no assurance for the success of cloud seeding operations, it is important to determine or forecast the success rate before any operations are conducted. Several climate factors, precipitation records and prediction results from the cloud models such as the Great Plains Cumulus Model (GPCM) are normally used in making the decision on whether the cloud seeding operation will be launched or not [7]. However, rainfall estimates are principal to evaluate the effectiveness of cloud seeding programs.

Traditionally, rainfall estimates have been mainly derived and forecasted from numerical modeling with both radar and ground observations [2,6,9]. As an alternative, this research presents the methodology from machine learning approaches to short-term rainfall forecasting. The goals of this paper are to present the methodology from machine learning approaches to short-term rainfall forecasting, including (1) how feature selection can be used to identify the relationships between rainfall occurrences and other weather conditions and (2) what models can be simply developed and deployed for predicting the accurate rainfall estimates to support the decisions in seeding operations.

II. Materials and Methods

Two integrated datasets, so-called GPCM and GPCM+RADAR, provided by Bureau of the Royal Rain Making and Agricultural Aviation and Department of Meteorology, Thailand, were explored in this study. The GPCM dataset consists of 309 daily records including the upper air observations, seeding operations and the average of rain volumes (AVG) from 18 rain gauges at regional weather stations. Each GPCM record contains 52 variables or features, for example, temperature, humidity, pressure, wind, atmospheric stability, seeding potential, operation and rain occurrence. The GPCM+RADAR dataset containing 179 records was made by linking the GPCM dataset with radar observations from the Chalermprakiat Royal Rainmaking Research center at Pimai, Nakhon Ratchasima Province during March 2004-September, 2006. More features

including the number of clouds, cloud base height, cloud intensity, and rain coverage area, were subsequently added, so each GPCM+RADAR record is 57 features in total. Based on the AVG feature, each record in both datasets was then categorized into (a) rain or no rain events, and (b) rainfall levels (no rain: 0-0.1 mm, few: 0.1-10 mm and moderate: >10 mm).

A two-step supervised learning framework were employed and evaluated using WEKA version 3.5.4 in model development for rainfall prediction in short-time period. In order to find the relationship among weather conditions and rainfall estimates, a correlated-based feature selection (CFS) was incorporated in the first step [4, 5] to filter out some noisy features and obtain the most important features for rainfall prediction. Applied with correlation, the CFS evaluates the relevance of features with respect to their ability to separate the classes. Next, the selected features from previous step were used to assess the performance of prediction models using three machine learning algorithms: C4.5 Decision Tree, Multilayer Perceptron Artificial Neural Network (MLP-ANN) and Support Vector Machine (SVM) [1,3,8].

III. Results

Using the GPCM and GPCM+RADAR datasets, the C4.5 decision-tree induction model can achieve accuracy of 87.06% and 94.41% respectively in forecasting whether rain or no-rain event, but provides somewhat lower accuracy at 62% level in forecasting whether no-rain, few-rain, or moderate-rain event will occur (Table I). However, the decision tree, as shown in Fig. 1 and Fig. 2, captures the structured decision making process that can be expressed as a rule set of *if-then statements* and easier for human to understand.

IV. Conclusion

It was found that all models in our study perform somewhat better for GPCM dataset than GPCM+RADAR dataset in both predictions of rainfall occurrence and classification of rain levels within same-day period. This indicated that too many or possibly redundant features can cause the rainfall forecasting to be inefficient and lower the accuracy. Therefore, the selection of relevant features and elimination of irrelevant and redundant ones are primarily need to increase in prediction accuracy and avoid over fitting of the training data. Results from the prediction on the weather stations level also showed that using only the data collected systematically and specifically for weather forecasting purposes at the local weather stations can improve the prediction accuracy.

In conclusion, the choice and the number of features selected to achieve the best performance in prediction of rainfall occurrences may vary by feature selection approaches, prediction algorithms, and the quality of training data. However, as evidenced in our results, the methodology from machine learning approaches can be used to facilitate monitoring of weather conditions and forecasting rainfall for a short-term period over the northeastern part of Thailand, and can apply to the conduct of appropriate seeding operations in other regions of Thailand.

Acknowledgment

The author thanks the Bureau of Royal Rainmaking and Agricultural Aviation (BRRAA) for providing fund in conducting this study. Particular thanks to these agencies: the Chalermprakiat Royal Rainmaking Center at Hua Hin, Prachuap Khiri khan, the Chalermprakiat Royal Rainmaking Research Center at Pimai, Nakhon Ratchasima and Thai Meteorological Department (TMD) as well as many individuals including the director and the administrators of the BRRAA, the director and the officers of the Royal Rainmaking Section of the BRRAA, and the director and the officers of the Chalermprakiat Royal Rainmaking Center at Hua Hin, the director of the Chalermprakiat Royal Rainmaking Research Center at Pimai, and the officers of TMD for invaluable advice and supplying data that the project could not accomplish without.

References

[1] M.J.A. Berry, and G.S. Linoff, Mastering Data Mining. New York: John Wiley & Sons, Inc, 2000.

[2] D.C. Blanchard, "Raindrop Size Distribution in Hawaiian Rains," Journal of Meteorology, vol. 10, pp. 457-473, 2000.

[3] A.J. Gasiewski, G.A. Showman, and G.M. Skofronick, "Application of Neural Nets to Rain Rate Retrieval from Simulated Multichannel Passive Microwave Imagery," In Geoscience and Remote Sensing Symposium. Remote Sensing for a Sustainable Future. IGARSS '96. International, vol. 3, pp. 1688 – 1691, 1996.

[4] M. A. Hall, and L. A. Smith, "Feature subset selection: a correlation based filter approach," International Conference on Neural Information Processing and Intelligent Information Systems. Springer, 1997, pp. 855-858

[5] M. A. Hall, and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper," Florida Artificial Intelligence Symposium. AAAI Press, 1999, pp. 235-239.

[6] J. Joss, K. Schran, J.C. Thoms, and A. Waldvogel, "On the Quantitative Determination of Precipitation by Radar" Wissenschaftclich Mitfeilung, No.63, Eidgennossischen Komnission Sum Studiumder Hagelgilbung and der Hergelshe, 1970.

[7] W. Khantiyanan, Analysis of GPCM forecasting model results. Bureau of Royal Rainmaking and Agricultural Aviation. Thailand, 1996.

[8] G.R. Kohavi, and John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, pp. 273-324, 1997.

[9] J.S. Marshall, and W. Mc. Palmer, "The Distribution of Raindrops with Size," Journal of Meteorology, vol 5, pp. 165-166, 1948.

TABLE I
THE OVERALL CLASSIFICATION ACCURACY OF THE DECISION-TREE MODELS IN PREDICTION OF RAINFALL OCCURRENCE WHEN USING THE GPCM AND THE GPCM + RADAR DATASETS WITH FEATURE SELECTIONS

Classification Accuracy of Rainfall Events	GPCM dataset	GPCM + RADAR dataset
rain/no rain	87.06% (13 features)	94.41% (13 features)
no rain/few-rain/moderate-rain	62.46% (11 features)	62.57% (9 features)

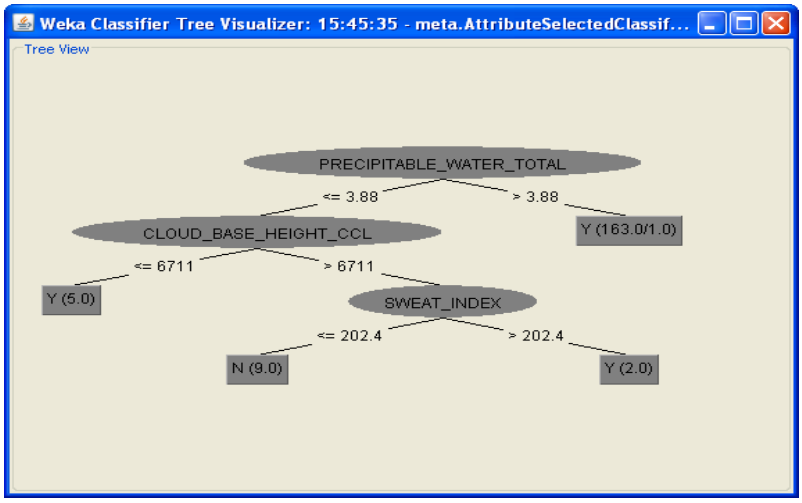

Fig. 1 Decision tree model for prediction of rainfall occurrence (rain/no rain) when using the GPCM dataset with feature selection

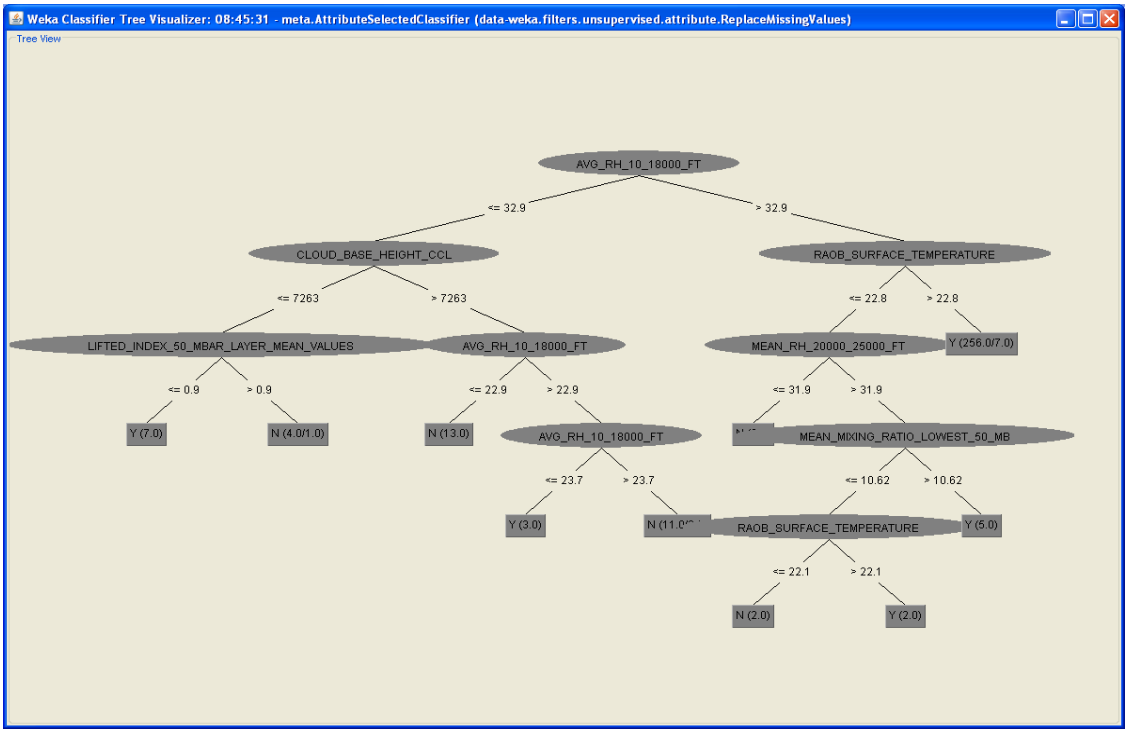

Fig. 2 Decision tree model for prediction of rainfall occurrence (rain/no rain) when using the GPCM+ RADAR dataset with feature selection

Comparing the MLP-ANN models at different time-points (same-day, next-day, next 2-day), the root means square error (RMSE) of the models for rainfall estimation in same-day are slightly lower than the models for both next-day and next 2-day estimates (Table II). The features selected for the MLP-ANN based rainfall estimation are also illustrated in Fig. 3 and Fig. 4.

TABLE II
THE ROOT MEANS SQUARE ERROR (RMSES) OF THE ARTIFICIAL NEURAL NETWORK MODELS FOR RAINFALL PREDICTION IN SHORT-TIME PERIOD WHEN USING THE GPCM AND THE GPCM + RADAR DATASETS WITH FEATURE SELECTIONS

MLP-ANN	GPCM dataset		GPCM + RADAR dataset	
	RMSE	# features	RMSE	# features
Same-day	0.155	7	0.171	5
Next-day	0.183	8	0.206	6
Next 2-day	0.169	13	0.207	14

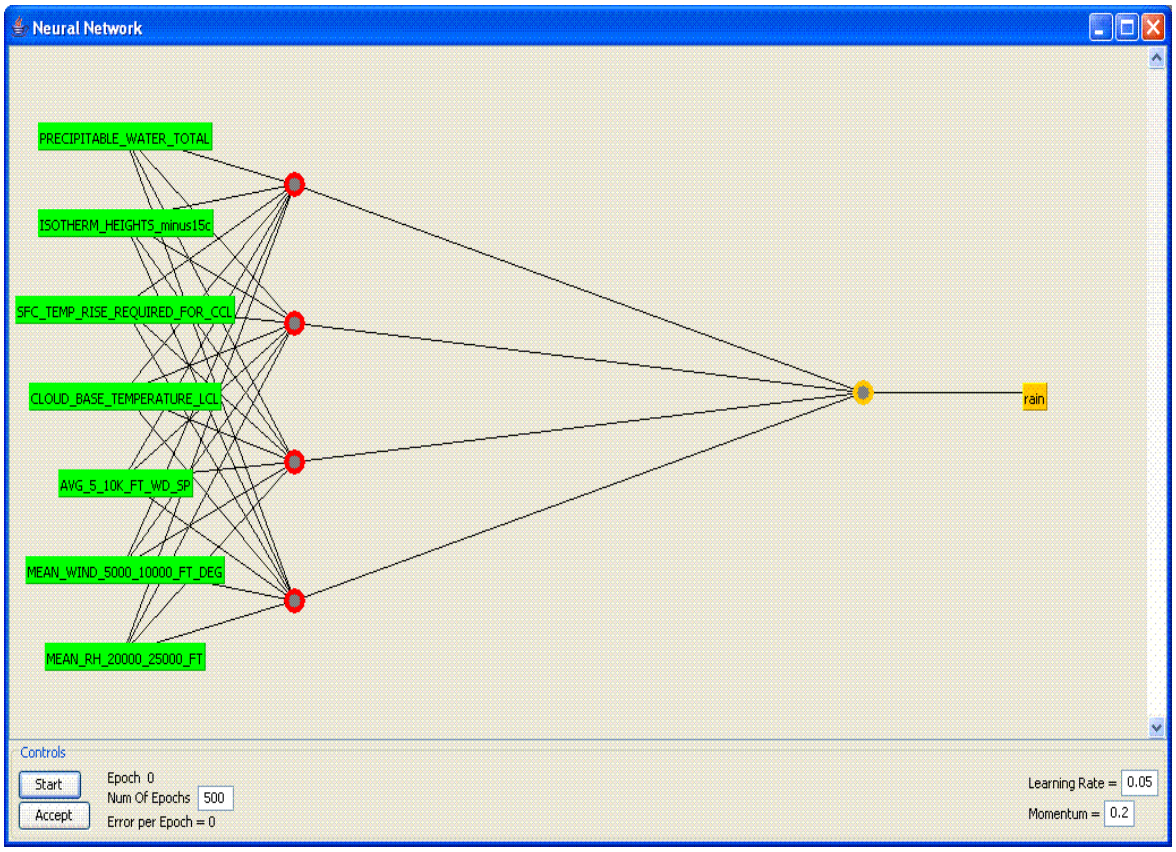

Fig. 3 Artificial neural network model for same–day prediction of rainfall estimates when using the GPCM dataset with feature selection

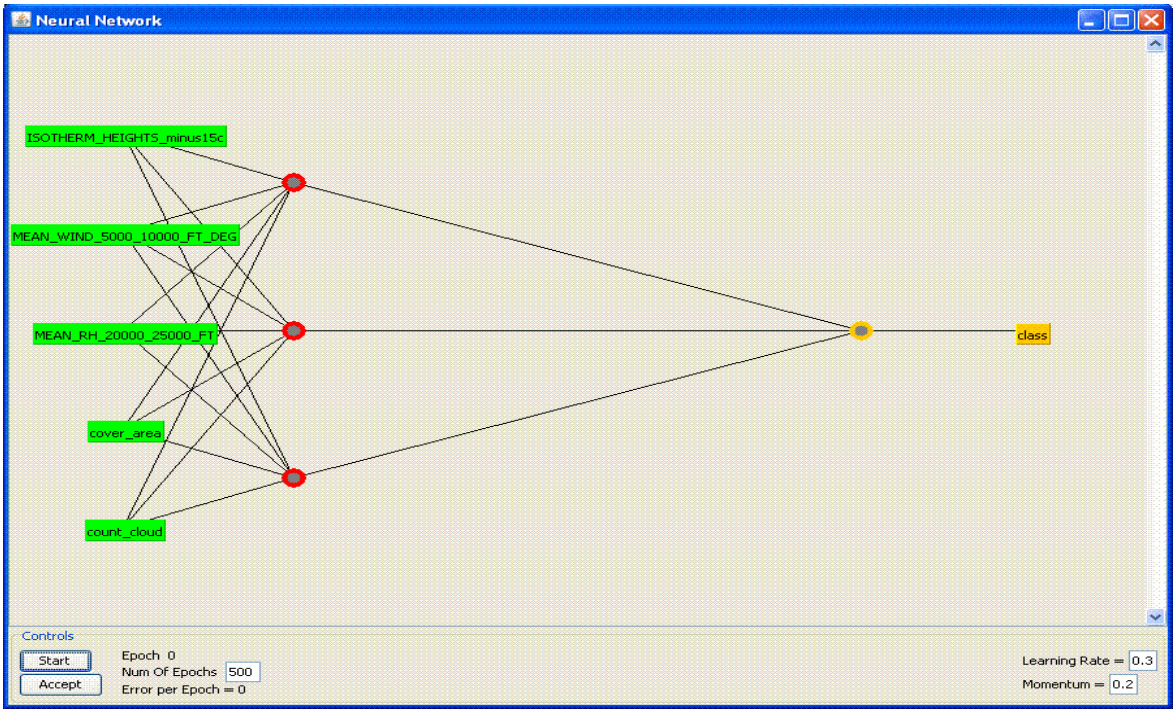

Fig. 4 Artificial neural network model for same-day prediction of rainfall estimates when using the GPCM+ RADAR dataset with feature selection

The MLP-ANN models for same-day prediction of rain levels yield the accuracy of 85.4% which are higher than the model performance for next-day and next 2-day predictions, on the GPCM dataset (Table III). Likewise, the SVM models for same-day prediction of rain levels perform better than the next-day and next 2-day predictions, on the GPCM dataset (Table IV). These results imply that the prediction models of rainfall estimates performed better or more accurate for short time periods.

TABLE III
THE OVERALL CLASSIFICATION ACCURACY OF THE MLP-ANN MODEL IN PREDICTION OF THREE RAINFALL LEVELS (NO RAIN, FEW, AND MODERATE) WHEN USING THE GPCM AND THE GPCM+RADAR DATASETS WITH FEATURE SELECTIONS

MLP-ANN	GPCM dataset		GPCM + RADAR dataset	
	accuracy (%)	# features	accuracy (%)	# features
Same-day	85.44	13	68.16	9
Next-day	81.95	7	58.60	6
Next 2-day	78.71	14	70.29	5

TABLE IV
THE OVERALL CLASSIFICATION ACCURACY OF SVM IN PREDICTION OF THREE RAINFALL LEVELS (NO RAIN, FEW, AND MODERATE) WHEN USING THE GPCM AND THE GPCM+RADAR DATASETS WITH FEATURE SELECTIONS

SVM	GPCM dataset		GPCM + RADAR dataset	
	accuracy (%)	# features	accuracy (%)	# features
Same-day	86.03	13	69.10	9
Next-day	82.97	7	64.10	6
Next 2-day	83.47	14	69.34	5

On the weather stations level, the prediction performances of the MLP-ANNs for rainfall estimates are shown in Table V and results from rain-level classification using Decision Tree and MLP-ANN algorithms are presented in Table VI. With the same testing criteria, the MLP-ANN models appear to provide superior prediction performance compared to the decision-tree inductions.

TABLE V
THE RMSES OF MLP-ANNS FOR RAINFALL PREDICTION IN SHORT-TIME PERIOD BY WEATHER STATIONS, WHEN USING THE GPCM AND THE GPCM+RADAR DATASETS WITH FEATURE SELECTIONS

Station	Same-day				Next-day				Next 2-day			
	GPCM		GPCM + RADAR		GPCM		GPCM + RADAR		GPCM		GPCM + RADAR	
	RMSE	# feature	RMSE	# feature	RMSE	# feature	RMSE	# feature	RMSE	# feature	RMSE	# feature
BURIRUM	0.100	6	0.117	6	0.104	7	0.114	7	0.116	5	0.112	5
CHAIYAPHUM	0.123	3	0.164	4	0.122	2	0.208	6	0.123	9	0.214	5
CHOK CHAI	0.161	7	0.187	4	0.166	7	0.181	3	0.156	9	0.204	4
KAMALASAI	0.088	6	0.102	4	0.092	11	0.122	7	0.094	14	0.119	9
KHON KAEN	0.090	9	0.114	7	0.103	5	0.121	3	0.097	6	0.133	6
KOSUM PHISAI	0.159	8	0.184	6	0.201	6	0.206	5	0.169	6	0.227	9
LOEI	0.136	8	0.159	4	0.155	3	0.169	6	0.133	14	0.182	7
MUKDAHAN	0.183	6	0.218	5	0.189	7	0.243	8	0.192	7	0.249	8
NAKHON PHANOM	0.129	4	0.162	6	0.147	7	0.175	9	0.129	10	0.205	13
NAKHON RATCHASIMA	0.099	6	0.118	4	0.100	6	0.120	7	0.167	7	0.155	9
NANG RONG	0.119	8	0.136	4	0.123	5	0.137	6	0.117	11	0.125	8
NONG KHAI	0.115	6	0.140	7	0.174	3	0.150	4	0.124	5	0.142	6
ROI ET	0.102	6	0.126	4	0.119	9	0.142	4	0.116	11	0.143	6
SAKON NAKHON	0.137	8	0.171	10	0.151	5	0.194	5	0.188	7	0.235	7
SURIN	0.090	9	0.113	6	0.112	11	0.123	9	0.103	13	0.137	12
THA TUM	0.151	8	0.168	3	0.146	3	0.163	5	0.144	8	0.155	8
UBON RATCHA THANI	0.158	9	0.199	10	0.175	11	0.187	4	0.150	7	0.182	7
UDON THANI	0.114	5	0.163	6	0.126	7	0.143	3	0.142	8	0.167	8

TABLE VI
THE ACCURACY CLASSIFICATION ASSESSMENT OF THREE RAINFALL LEVELS (NO RAIN, FEW, AND MODERATE) BY WEATHER STATIONS, WHEN USING THE GPCM AND THE GPCM+RADAR DATASETS WITH FEATURE SELECTIONS

Station	Artificial Neural Network Model		Decision Tree Model	
	GPCM	GPCM + RADAR	GPCM	GPCM + RADAR
BURIRUM	58.58	56.98	58.90	53.07
CHAIYAPHUM	63.11	55.87	58.90	55.87
CHOK CHAI	62.14	56.98	57.61	50.28
KAMALASAI	62.14	55.31	56.31	48.60
KHON KAEN	62.78	55.31	63.43	55.87
KOSUM PHISAI	61.17	51.96	60.84	47.49
LOEI	56.96	53.07	50.49	53.63
MUKDAHAN	60.84	49.72	57.93	43.58
NAKHON PHANOM	62.46	58.66	60.19	52.51
NAKHON RATCHASIMA	62.14	64.25	60.84	57.54
NANG RONG	63.75	54.75	64.72	53.63
NONG KHAI	61.81	50.28	57.61	47.49
ROI ET	58.90	49.72	55.66	48.60
SAKON NAKHON	53.40	41.90	53.07	45.81
SURIN	62.14	55.31	51.13	52.51
THA TUM	61.49	59.22	57.28	48.04
UBON RATCHA THANI	58.58	51.96	55.99	48.04
UDON THANI	58.25	51.40	53.72	48.60