

Received January 7, 2020, accepted January 21, 2020, date of publication February 10, 2020, date of current version February 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972435

Annual and Non-Monsoon Rainfall Prediction Modelling Using SVR-MLP: An Empirical Study From Odisha

XIAOBO ZHANG^{ID1}, SACHI NANDAN MOHANTY^{ID2}, AJAYA KUMAR PARIDA^{ID3},
SUBHENDU KUMAR PANI^{ID4}, BIN DONG⁵, AND XIAOCHUN CHENG⁶

¹Faculty of Automation, Guangdong University of Technology, Guangzhou 510006, China

²Department of Computer Science and Engineering, Faculty of Science and Technology (FST), ICFAI Foundation for Higher Education, Hyderabad 501203, India

³Department of Computer Science and Engineering, KIIT University, Bhubaneswar 751024, India

⁴Department of Computer Science and Engineering, Orissa Engineering College, Bhubaneswar 752050, India

⁵Affiliated Hospital of Hebei University, Hebei University, Baoding 071000, China

⁶Department of Computer Science, School of Science and Technology, Middlesex University, London NW4 4BT, U.K.

Corresponding author: Bin Dong (dongbin@hbu.edu.cn)

ABSTRACT Rainfall is a natural demolishing phenomenon. On the other side, it also serves as a major source of water when conserved through proper channel. For this issue, estimation of rain fall is of utmost importance. The present study employed on rain fall forecasting in annual as well as non-monsoon session in Odisha (India). The total annual rainfall and relative humidity data were collected from period 1991-2015 from Department of Forest and Environment Govt. of Odisha. Support Vector Regression and Multilayer perception implemented for prediction of maximum rainfall in annual and non-monsoon session. Input parameter like average temperature in month, wind velocity, humidity, and cloud cover was considered for predicting rainfall in non-monsoon session. The performance of the results was measured with MSE (mean squared error), correlation coefficient, coefficient of efficiency and MAE (mean absolute error). The results of SVR were compared to those of MLP and simple regression technique. MLP being a computationally intensive method, SVR could be used as an efficient alternative for runoff and sediment yield prediction under comparable accuracy in predictions. SVR-MLP may be used as promising alternative forecasting tool for higher accuracy in forecasting and better generalization ability.

INDEX TERMS Support vector regression (SVR), multi-layer perception, rainfall.

I. INTRODUCTION

Rainfall Prediction is the process of predicting the probability of precipitation in a particular region and forecasting of rainfall in future along with estimating the amount of rainfall in specific regions. It takes into account both accuracy of prediction, error in prediction and rainfall volume estimation along with probability of rainfall in that specific region. It is prepared by forecasters by way of collecting, analyzing, verifying, modeling, simulating and doing research on different meteorological data and parameters available. Some basic parameters include Average Mean Monthly Temperature (Minimum and Maximum), Total annual rainfall in mm and Relative Humidity. Understanding the probability of rainfall along with its intensity (heavy/light) helps farmers as well as city population, being alert toward any

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson^{ID}.

disaster which can cause damage to life and property. Traditional statistical models are widely used for rainfall forecasting. But it is observed by many researchers that the inclusion of machine learning techniques for rainfall prediction and forecasting has been used for high accuracy in general as well as complex situations when dealing with peak values of rainfall. The major challenge is handling highly non-linear data and low peak and high rainfall values. Hence machine learning techniques such as multilayer perceptron (MLP) and Support Vector Machines (SVM) [2] are used by researchers to handle these challenges. In this paper two separate regression processes have been implemented: one for annual and the other for 8 months non-monsoon session. Regression analysis has been performed using 2 machine learning techniques, one is SVM and the other is MLP in each case. In this study 24 years past data has been collected which may act as a major contribution of the study.

The motivational factor in this study is to traditional optimization methods, which emphasize accurate and exact computation for the global optimum of a continuous function and avoid being trapped into one of the local optima, but may fall down on achieving the global optimum. In this study SVR-MLP implemented for achieving global optima.

II. BACKGROUND

A. REGRESSION

Regression is a type of supervised learning method which is used for prediction in discrete amount rather than a particular class prediction. It is to be used when we want to predict concrete values as output. Here, basically a relationship has to be established between the dependent variable (say Y) with independent variable (say X). This relationship may be used to predict future values in predictive analysis.

Simple Linear Regression is given by $y = b + b_1 x$, where y is the dependent variable whose value is to be predicted and x is the independent variable whose value is known. b is constant value and b_1 is regression coefficient.

Multiple Regressions is an extension of simple linear regression. It has been given as is given as $y = f(x_1, x_2, \dots, x_n)$, where y is the single dependent variable and x_1, x_2, \dots, x_n are multiple independent variables.

Multivariate Regression deals with multiple dependent variables and multiple independent variables. In this paper multiple regressions have been carried out.

B. SUPPORT VECTOR REGRESSION OR SUPPORT VECTOR MACHINES (SVM) FOR REGRESSION

The use of SVM for regression is known as Support Vector Regression (SVR) [3]. This algorithm automatically converts nominal values to numeric values. Input data set has to be normalized before training start, either automatically (by tool setup or scripting) or manually by the user (data-set normalization). SVR finds a best fit line which reduces the error of the cost function. Only those instances (Support Vectors) in the training data set are chosen which are nearest to the line with minimum cost. A margin has to be created around the line for better adjustment of prediction and then the data may be projected into higher dimensional space for better prediction and flexibility.

The cost function minimizes the parameters over the dataset. It is defined as: $J(\theta) = 1/2(||\theta||^2)$. The main objective is to minimize θ to produce optimized result.

Assuming that the data has two features X and Y , a support vector machine model is shown below in figure 4.

Kernel Functions are used to handle the high dimensionality of the feature space. Proper selection of Kernel function can produce more effective result or accuracy in least time thus increasing efficiency of the model. Weka tool uses various kernels to achieve this task [2]:

i.) **Linear Kernel:** Here, data is separated by straight line or hyperplane.

ii.) **Polynomial Kernel:** This SVM kernel can handle non-linear models by populating the vectors in a feature space over polynomials of the original variables. It uses interaction features by combining features of input data samples and then determines their relationship. The *Poly* kernel can be given as: $K(x, w) = (k + x^T w)^d$, where x and w are vectors in input space from training or testing data and T is a constant trade off parameter.

iii.) **RBF Kernel:** The Gaussian Radial basis Function kernel is defined by $F(x_i, x_j) = \exp(-||x_i - x_j||^2 / 2\sigma^2)$, where x_i and x_j are two feature vectors in some input space and $||x_i - x_j||$ is the Euclidean distance between the two feature vectors and σ is a free parameter.

iv.) **PUK Kernel:** In Weka PUK Kernel is Pearson VII function based Universal Kernel. The Pearson VII function was developed in 1895 by Karl Pearson. In this paper PUK Kernel is utilised. It serves as a Universal kernel if parameters are adjusted accordingly. The PUK kernel is defined as: $F(x_i, x_j) = 1 / [1 + \{2^* \sqrt{(||x_i - x_j||)^2 * \sqrt{2^{(1/\Omega)} - 1}}\} / \sigma^2]^\Omega$ where (x_i, x_j) are the two vectors forming multi-dimensional input space, $||x_i - x_j||$ is the Euclidean distance and σ and Ω are the width controlling parameters and are adjusted for proper curve fitting. In this paper PUK Kernel using Weka has been implemented.

C. MULTI-LAYER PERCEPTION

A more efficient Feed Forward Back Propagation network is a model that is trained using a back-propagation training algorithm. The back-propagation training algorithm finds the error amount by subtracting the training output from the desired output to obtain the error amount. The error is then back propagated and weights are updated accordingly. The weights and biases in the hidden layers till the input layer are adjusted or updated to reduce the error amount. A Multi-Layer Perception (MLP) [5] network is a type of Feed Forward Back Propagation neural network which is trained using a back-propagation algorithm. A MLP consists of minimum three layers: a single input layer, at least one hidden layer and single output layer. There can be more than one number of hidden layers. One can also conclude that Deep Neural Network (DNN) is a sub-category MLP of since a deep neural network is a type of Artificial Neural Network that can easily handle a huge number of hidden layers efficiently. This model is highly efficient in handling non-linearity and peak values of data and also for producing an accurate future prediction [21], [22]. So basically Multi-Layer Perception and Deep Neural Networks are almost similar models that use back-propagation except for the fact that DNN are deeper models than MLP which means that a DNN can handle up to thousands of hidden layers accurately and efficiently while in MLP we limit the hidden layers to fewer numbers as compared to DNN (but multiple layers). MLP can study data that is not linearly separable. In Weka, the nodes in this network are all sigmoid (except for when the class is numeric, in which case the output nodes become unthreshold linear units).

Sigmoid Activation Function: This is in MLP given by $\sigma(\mathbf{z}) = 1/(1+e^{-\mathbf{z}})$.

The basic steps in MLP training are:

- i.) Initializing the weights.
- ii.) Computing error function using error back-propagation algorithm and learning rate (η).
- iii.) Updating the estimated weight by conjugate gradient Optimization.

One set of updates for all the weights and the training patterns is called one training epoch for output, the weight change is given by $\Delta w_{kj} = \eta \cdot \delta_k \cdot o_j$, where δ_k is the output error.

Conjugate Gradient Optimization: The conjugate gradient algorithm is a modified version of steepest descent based on a theoretical foundation. It is more efficient and reliable than the back-propagation algorithm and its variations. Conjugate gradient does not have any heuristic factors or matrix computation so it is more efficient. It is used to choose a suitable search direction during search minimization. The new search direction at the new minimum point is found by setting $\mathbf{P}_{k+1} = -\mathbf{g}_{k+1} + \beta_{k+1} \mathbf{P}_k$, where \mathbf{P} is the search direction at different points.

III. RELATED WORKS

Samsudin et al. (2010) tested the flexibility of the SVM in time series forecasting and compared it with a multi-layer back-propagation (BP) neural network. SVM outperformed the multi-layer back-propagation neural network. The Mean Absolute Error (MAE) was used as comparison parameter. SVM proved to be an efficient technique in time series forecasting techniques.

Trivedi and Deny (2013) described the effects of interaction among various Kernels, Normalized Polynomial Kernel (NP), Polynomial Kernel (PK), Radial Basis Function Kernel (RBF), and Pearson VII Function based Universal Kernel (PUK) with three feature selection techniques such as Gain Ratio (GR), Chi-Squared and Latent Semantic Indexing (LSI) have been implemented on the Enron Email Data Set. Polynomial Kernel (NP) performs the best for all the tested feature selection techniques PUK kernel became the second best in performance and performed well with low dimensional data but shows poor performance for high dimensional data Latent Semantic Indexing (LSI) performed the best amongst all the tested feature selection techniques. Out of all the kernel functions, Polynomial kernel performs best for all the tested features. PUK kernel became the second best in performance and performed well with low dimensional data but shows poor performance for high dimensional data. Latent Semantic Indexing (LSI) performed the best amongst all the tested feature selection techniques.

Koskela et al. (1996) compared Multilayer perception network (MLP), FIR neural network and Elman neural network in four different time series prediction tasks. They showed that the efficiency of the learning algorithm is more crucial factor than the model used. Elman network models showed better results than MLP in an electric network series and

Elman network showed similar prediction results like MLP in other prediction tasks. FIR network showed satisfactory performance but not as good results as Elman network.

Ustun et al. (2006) used Pearson VII function(PEK) for Support Vector Regression(SVR) and compared its performance with other commonly applied SVM kernels by applying it to simulated as well as real world data sets. It was concluded that PUK Kernel was more robust and provides better mapping power than other SVM Kernels.PEK provides better generalization performance of SVM and can be used as a Universal Kernel to serve as an alternative to linear, polynomial and RBF kernels.

Kumarasiri and Sonnadara (2006) utilized Artificial Neural Networks based on feed-forward back-propagation architecture for rainfall forecasting. Their main aim was to make predictions from the available data, not on focusing more on the physical aspects of the atmosphere or the actual process of rainfall occurrence. They attempted both short term and long term forecasting for ground level collected data from meteorological verified sources in Colombo, SriLanka. They developed three Neural Network models for rainfall prediction and the success rate and trends of rainfall for monsoon season were analyzed by them.

Chattopadhyay and Chattopadhyay (2007) used multi-layer perceptron technique to predict the average monsoon rainfall over India. They used the data for three months June, July, and August which is considered as summer-monsoon months from period 1871 to 1999.Only one parameter which is historical rainfall data was used as input parameter to predict future average summer-monsoon rainfall. This paper develops ANN model step-by step to predict the average rainfall over India. Performance of the model is evaluated through computation of overall prediction error (PE) They concluded that Soft Computing as Artificial Neural Network can be of great use in prediction monsoon rainfall over India. If more input parameters are available, then a prediction of higher accuracy would be possible.

Paras et al. (2009) used feature based forecasting and time series using ANN to predict various weather parameters like maximum temperature, minimum temperature and relative humidity by using feature extraction as well as the parameter time series data.The feed forward artificial neural networks with back propagation was trained to predict future weather conditions using collected data of particular station.

Hung et al. (2009) presented a new approach using an Artificial Neural Network technique to improve performance of rainfall forecasting. 4 years of hourly data from 75 rain gauge stations in the area were used to develop the ANN model. The developed ANN model was used for real time rainfall forecasting and flood management in Bangkok, Thailand.

Monira et al. (2010) modeled ensemble methods using Logit Boosting (LB), and Random Forest (RF). They also modeled a single classifier model using Least Square Support Vector Machine (LS-SVM). They optimized each of the models on validation sets and then forecasted with the optimum model on the test dataset. The results suggested that these

methods are capable of efficiently forecasting the categorical rainfall amount in short term.

Deshpande (2012) uses multi-layer perceptron to predict time series data. Rainfall samples were collected from the authorized Government Rainfall monitoring agency in Yavatmal, Maharashtra, India. Only one parameter that is historical monthly rainfall data is used to predict future rainfall. Multi-step ahead (1, 5, 10, 20) predictions of this Rainfall Data series was carried out using the proposed Multilayer Perception Neural Network. The comparison and analysis of Jordon Elman and MLP network were carried out. The multi-layer perception successfully predicted rainfall time series much better than Jordon Elman network. The result was calculated by using software, “Neurosolution 5.0”. Only one parameter that is historical rainfall data was used.

Abhishek et al. (2012) presented prediction possibility of average rainfall in Udupi district of Karnataka by using artificial neural network models. In formulating these predictive models they developed three layered network. They used back-propagation-feed forward neural network for finding the number of hidden neurons in these layers for the best performance model.

Sumi et al. (2012) modeled a hybrid multi-model method and compared it with its constituent models. The models include the artificial neural network, multivariate adaptive regression splines, the k-nearest neighbor, and radial basis support vector regression. Each of these methods is applied to model the daily and monthly rainfall, coupled with a pre-processing technique including moving average and principal component analysis. In the first stage of the hybrid method, sub-models from each of the above methods are constructed with different parameter settings. In the second stage, the sub-models are ranked with a variable selection technique and the higher ranked models are selected based on the leave-one-out cross-validation error. The forecasting of the hybrid model is performed by the weighted combination of the finally selected models. They concluded that hybrid model gives more accurate forecast than individual models for the daily rainfall series.

Mislam et al. (2015) applied Artificial Neural Network (ANN) with the Back propagation Neural Network (BPNN) algorithm. They tested the rainfall data using two-hidden layers of BPNN architectures with three different epochs. The mean square error (MSE) was used to determine performance of classification. The experimental results showed that BPNN algorithm is a good model to predict rainfall in Tenggarong, East Kalimantan - Indonesia.

Dubey (2015) created different artificial neural networks for the rainfall prediction of Pondicherry, a coastal region in India. The ANN models were created using three different training algorithms which are feed-forward back propagation algorithm, layer recurrent algorithm and feed forward distributed time delay algorithm. The mean squared error (MSE) was found out for each model and the best accuracy got was for feed-forward distributed time delay algorithm.

Sharma et al. (2015) compared the accuracy of SVM, ANN and regression models. SVM gave better accuracy than regression but satisfactory when compared to ANN. Daily rainfall and daily temperature data was used for 1995 to 1999 for runoff prediction and for 2001–2003 of the wet season for sediment yield prediction, using support vector machines (SVMs). The performance of the model was evaluated using the root mean square error, correlation coefficient and coefficient of efficiency. The results of SVM were compared to those of ANN and simple regression. They concluded that ANN being a computationally intensive method, SVM could be used as an efficient alternative for runoff and sediment yield predictions under comparable accuracy in predictions.

Lee et al. (2017) implemented the spatial prediction of flood susceptibility by using random-forest and boosted-tree models in Seoul metropolitan city, Korea. For the flood-susceptibility mapping, flooded-area, topography, geology, soil and land-use, they collected the datasets and entered them into spatial datasets. For training they used the flooded area of 2010 and used the flooded area of 2011 for the validation. The importance of the factors of the flood-susceptibility maps was calculated and lastly, the maps were validated. The random-forest model showed validation accuracies of 78.78% and 79.18% for the regression and classification algorithms, respectively, and boosted-tree model showed validation accuracies of 77.55% and 77.26% for the regression and classification algorithms, respectively. The flood-susceptibility maps provide meaningful information for decision-makers regarding the identification of priority areas for flood-mitigation management.

Das et al. (2017) tested a random forest based machine learning algorithm for forecasting of convective rain with a ground based radiometer. They used Brightness temperatures measured at 14 frequencies (7 frequencies in 22–31 GHz band and 7 frequencies in 51–58 GHz bands) as the inputs of the model. Synthetic minority over-sampling technique is used to balance the data set and 10-fold cross validation is used to assess the performance of the model. Results indicate that random forest algorithm with fixed alarm generation time of 30 minutes and 60 minutes performed quite well (probability of detection of all types of weather condition 90%) with low false alarms. Study shows the suitability of a random forest algorithm for forecasting application utilizing a large number of input parameters from diverse sources and can be utilized for other forecasting problems.

Supriya et al. (2015) used regression analysis using Annual maximum rainfall, stream flow and area of catchment to predict flood priority in different catchment. They found out that the Lower Vellar was more prone to flood and needed more control measures.

Zaman (2018) & Liu., Q, Zou (2019) developed two regression analysis and three classification analysis models for rainfall prediction of 33 Bangladeshi weather station. He used Apache Spark library for machine learning in scala programming language. His main idea behind the use of classification and regression analysis was to see the

comparative difference between types of algorithms prediction output along with usability. The study is further extended by developing another popular regression analysis algorithm named Random Forest Regression. After then, few other classification algorithms have been used for model building, training and prediction. Those are Naive Bayes Classification, Decision Tree Classification and Random Forest Classification. The developed and trained model is capable of predicting rainfall in advance for a month of a given year for a given area models are improved with the rainfall prediction accuracy.

IV. STUDY AREA

Odisha state lies on the eastern side of India adjacent to the Bay of Bengal. Odisha lies between 17.49N to 22.34N latitude and from 81.27E to 87.29E longitude. It is observed that in the monsoon period (June-September) the rainfall or flood can be somehow predicted properly in Odisha. But the rainfall amount prediction in non-monsoon periods (mainly October-November) is very unpredictable in Odisha.

From the past weather data, it can be verified that Odisha is under the influence of severe cyclonic storms. Sudden heavy floods during October-November have been caused huge damage to agriculture and property in villages as well as cities. Also, this unnatural cyclonic situation is mainly experienced in coastal Odisha while at the same time the interior areas of Odisha far away from coast experience draught situation. Thus non-monsoon period rainfall in Odisha is very erratic and is having Peak high and Peak low values in different regions simultaneously.

According to monsoon variations the climate of Odisha can be categorized into four types:

- 1) Pre-monsoon season (March – May)
- 2) South-west monsoon season or monsoon season (June-September)
- 3) Post-monsoon season (October-November)
- 4) Winter season (December-February)

In this study the period June-September (4 months) as mentioned above is considered as the normal monsoon season since this is the period of major rainfall in Odisha. The period from January-May (5 months) and October-December (3 months) which is a total of 8 months is considered as the non-monsoon rainfall period.

According to physiography, Odisha can be categorized into the following regions:

- 1) The eastern coastal plain.
- 2) The region of middle mountains and highlands.
- 3) The central plateaus.
- 4) The lower western rolling uplands.

V. DATA COLLECTION

The primary data has been collected for this study, data were collected from 1991-2005 across the Odisha state, consisting of parameters Average Monthly Temperature (Maximum), Average Monthly Temperature(Minimum), Relative Humidity and annual rainfall(mm.) was collected from Centre for

Environmental Studies (CES), from the Department of Forest and Environment, Government of Odisha.

For non-monsoon regression analysis monthly weather parameters Average Temperature (in degree Celsius), Cloud cover (percentage), Potential Evapotranspiration and rainfall (in mm.) data for Jagatsinghpur district of Odisha from 1960-2002 has been collected from the website <https://www.indiawaterportal.org/>.

VI. METHODOLOGY

Different methods have been adopted for analysis the data, by using Multiple Regression Analysis for prediction of annual rainfall the following steps are to be adopted:

- i. The input data set containing 4 weather parameters is observed carefully. All parameters are verified, and there is not missing value.
- ii. The multi-weather parameter data set used as input data is then normalized by Min-Max normalization between the value of 0 and 1 by using the formula:
$$Y_i = [X_i - \min(X)] / [\max(X) - \min(X)]$$
- iii. The normalized data set is used in Weka3.8.3 data mining tool where the normalized data set is split into 80% Train and 20% Test.
- iv. The above process is applied for both SVM (PUK Kernel) and MLP models. For SVM in Weka “SMOreg” option is selected and then kernel is selected as PUK. The result of regression in form of Correlation coefficient, Mean absolute error and Root mean squared error is visualised and noted. The regression vector for SVR is observed.
- v. For MLP regression, 2 hidden layers are chosen having 3 units each in form (3,3). The learning rate and momentum is set properly and error per epoch is calculated. All other result parameters i.e. Correlation coefficient(r), Mean absolute error and Root mean squared error are noted down. Weka 3.8 software is used for multiple regression analysis.
- vi. Both the models parameters, Correlation coefficient(r) mean absolute error and Root mean squared error, are compared to find the best rainfall prediction model for annual rainfall in Odisha. The multiple regression plots were analysed.

For Non-monsoon multiple regression analysis for Odisha for individual monthly rainfall prediction following steps are adopted:

- i) The collected data from 1960-2002 for Jagatsinghpur district of Odisha will be used to create model for entire state of Odisha. Since machine learning models depend on past data to provide predictions it is immaterial whether the data is state data or district data as long as the data parameters or attributes are constant. Moreover, a state data is average of all districts data only. Hence, if a model shows good performance for data of Jagatsinghpur district of Odisha then definitely it will show equivalently good performance for entire state

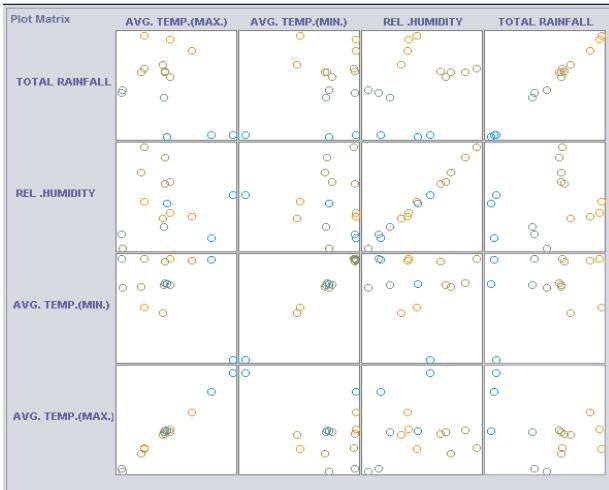


FIGURE 1. Result of annual rainfall prediction using SVM regression.

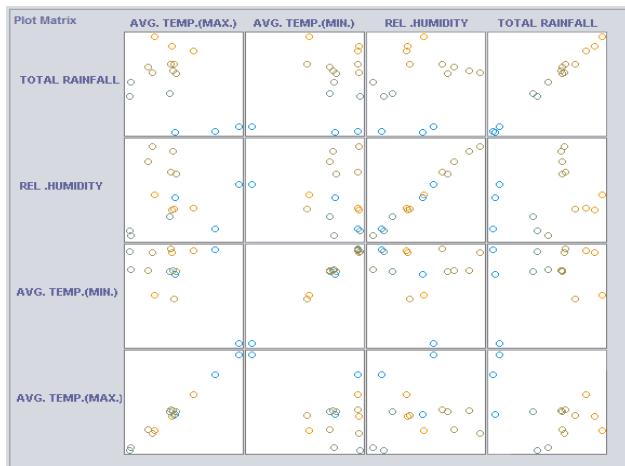


FIGURE 2. Result of annual rainfall prediction using MLP.

of Odisha. Hence, the machine learning model developed by using data of Jagatsinghpur district of Odisha using multiple regressions will serve as a standard non-monsoon rainfall prediction model for entire Odisha state.

ii) The collected data set from 1960-2002 containing 4 weather parameters is transformed to a different form. The transformation of data is done in following way:

a) 43 years of only January parameters (*Average Temperature, Cloud cover, Potential Evapotranspiration and rainfall data*) are kept in a separate csv file which is to be given as input to the model. The January data set is normalized by Min-Max normalization between value of 0 and 1 by the formula:

$$Y_i = [X_i - \min(X)] / [\max(X) - \min(X)]$$

This is the input for January regression analysis.

b) Similar process is carried out for rest 7 non-monsoon months. (February-May and October-December) using 1960-2002 data. So all total 8 different normalized csv files are created as 8 different inputs SVM and MLP each.

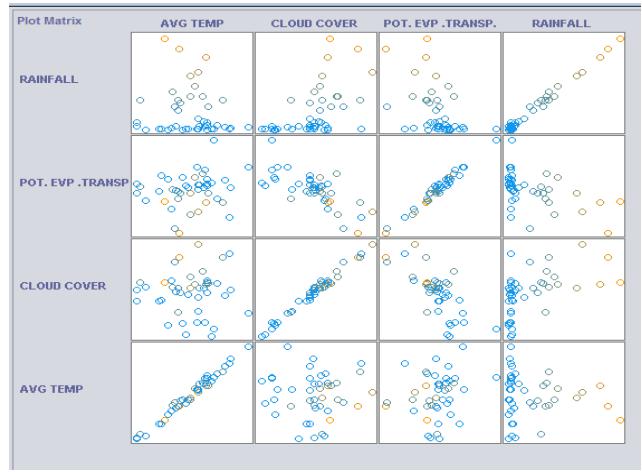


FIGURE 3. Result of rainfall prediction for January using SVM.

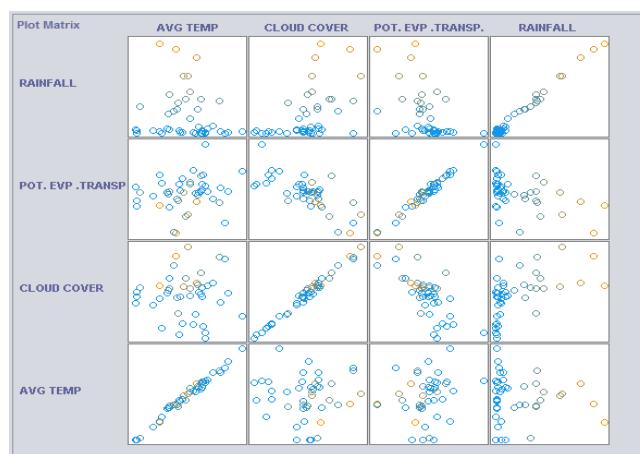


FIGURE 4. Result of rainfall prediction for January using MLP.

c) For individual months separate monthly multiple regressions is carried out using SVM and MLP both. Then a comparison is done based on the performance of both models.

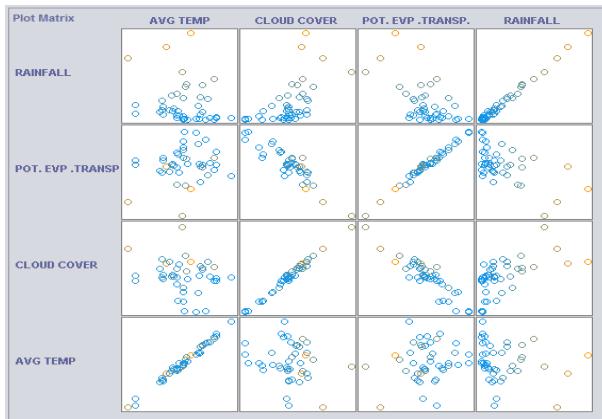
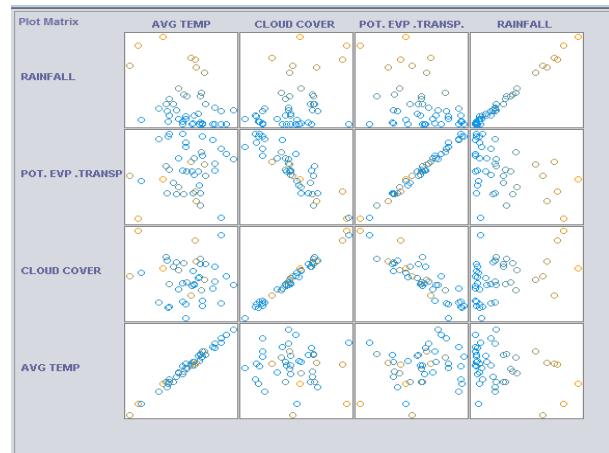
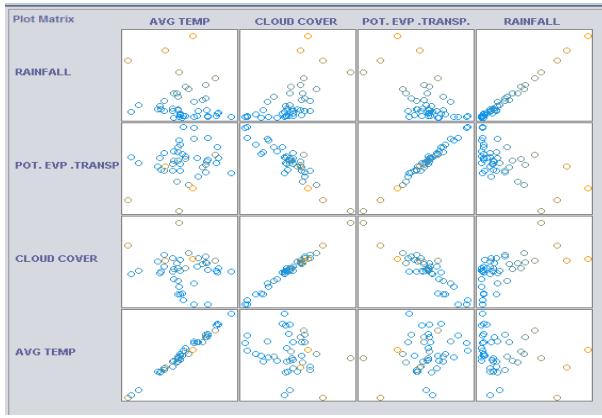
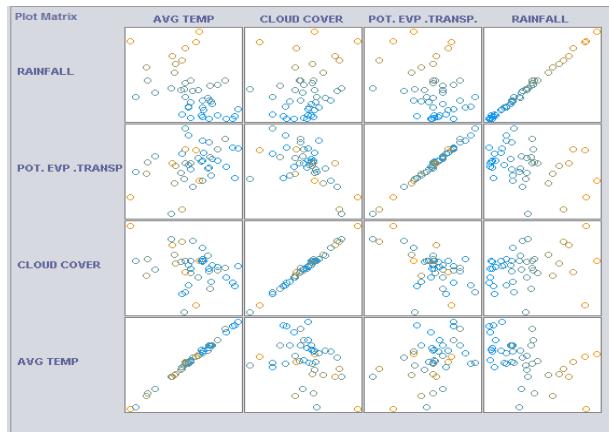
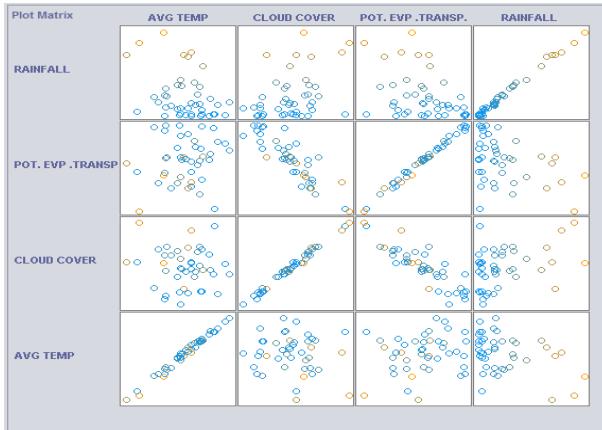
d) The percentage value of test-train split, choice of SVM Kernel, choice of MLP nodes and layers varies from month to month and is not constant. Weka data mining software is used for multiple regression analysis.

e) The comparison between the models are carried out by comparing values of Correlation coefficient(r), Mean absolute error (MAE) and Root mean squared error (RMSE). The multiple regression plots were created and analyzed using Weka data mining tool.

VII. RESULTS FOR MODEL USING SVM REGRESSION AND MLP

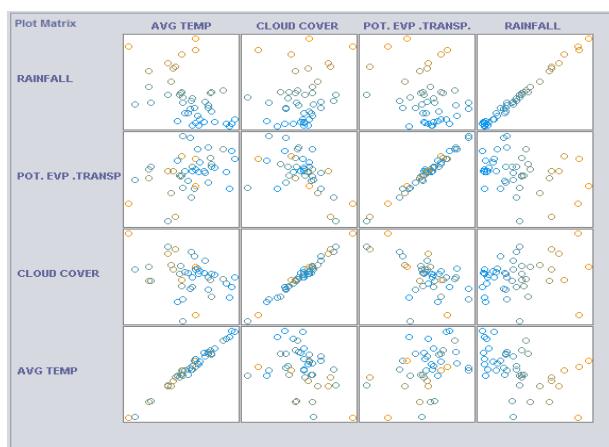
A. ANNUAL RAINFALL PREDICTION USING FOR ODISHA USING MULTIPLE REGRESSION ANALYSIS BY SVM AND MLP

The result of multiple regression analysis using both SVR and MLP using annual data for Odisha from 1991-2015 by using parameters, Average Monthly Temperature

**FIGURE 5.** Result of rainfall prediction for February using SVM.**FIGURE 8.** Result of rainfall prediction for March using MLP.**FIGURE 6.** Result of rainfall prediction for February using MLP.**FIGURE 9.** Result of April rainfall prediction using SVM.**FIGURE 7.** Result of rainfall prediction for March using SVM.

(Maximum), Average Monthly Temperature (Minimum), Total rainfall(mm.) and Relative Humidity was shown below.

Eighteen different models have been discussed with the help of SVM and MLP. Out of sixteen models developed for non-monsoon session, and two for annual rainfall prediction. As are shown in figure 1 and figure 2.

**FIGURE 10.** Result of April rainfall prediction using MLP.

B. PREDICTION OF NON-MONSOON USING MULTIPLE REGRESSIONS IN MONTHLY WISE

1) MONTH 1(JANUARY):NON-MONSOON REGRESSION ANALYSIS

Result of rainfall prediction using regression analysis with SVM and MLP kernels. As per shown in figure 3 and figure 4.

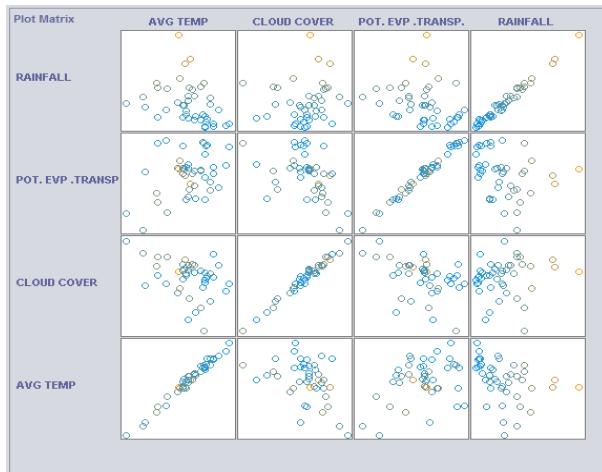


FIGURE 11. Result of May rainfall prediction by SVM.

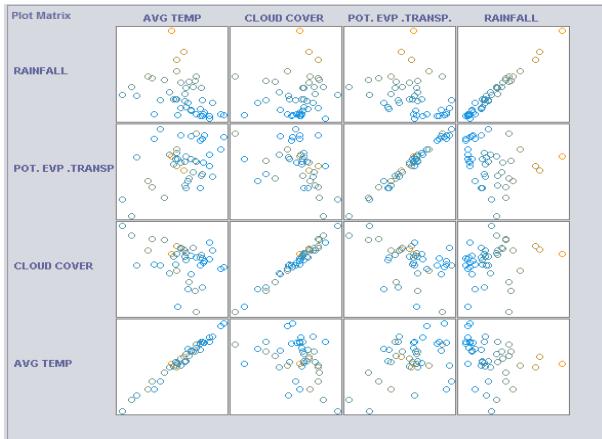


FIGURE 12. Result of May rainfall prediction by MLP.

2) MONTH 2(FEBRUARY):NON-MONSOON REGRESSION ANALYSIS

The result of non-monsoon rainfall prediction for February by multiple regression analysis using SVM RBF kernel and MLP (3, 3) is shown in figures below. The data set was split by 70% train and 30% test for both models.

3) MONTH 3(MARCH):NON-MONSOON REGRESSION ANALYSIS

The performance results of both SVM (Poly Kernel) model and MLP (3, 3) model for March rainfall prediction using multiple regression analysis is shown in figures below.

4) MONTH 4(APRIL):NON-MONSOON REGRESSION ANALYSIS

The performance results of both SVM (Poly Kernel) model and MLP (2, 2) model for April rainfall prediction using multiple regression analysis is shown in figures below. The data set was split by 70% train and 30% test for both models.

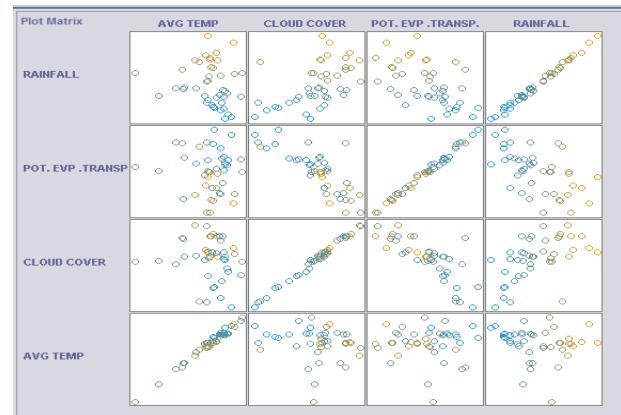


FIGURE 13. Result of October rainfall prediction by SVM.

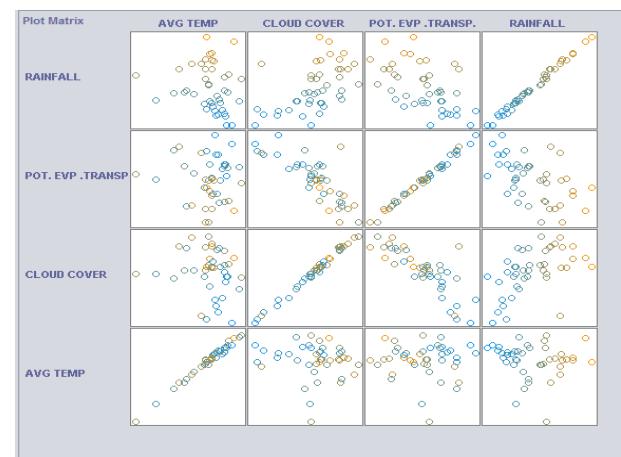


FIGURE 14. Result of October rainfall prediction by MLP.

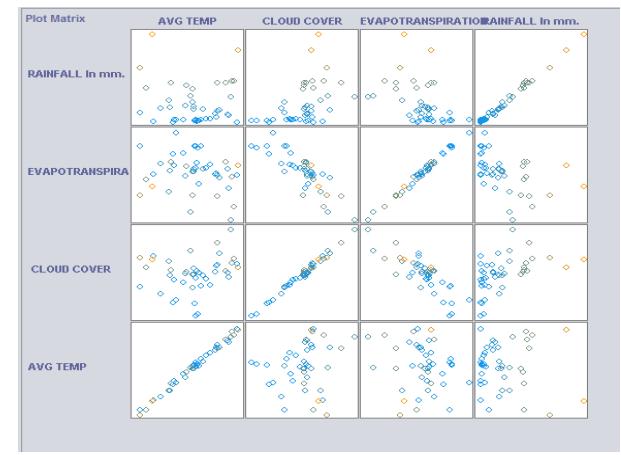
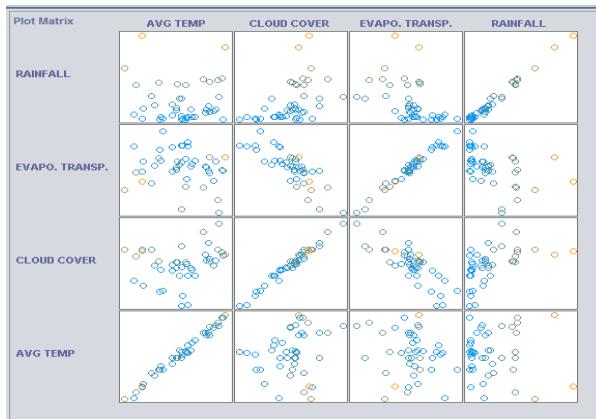
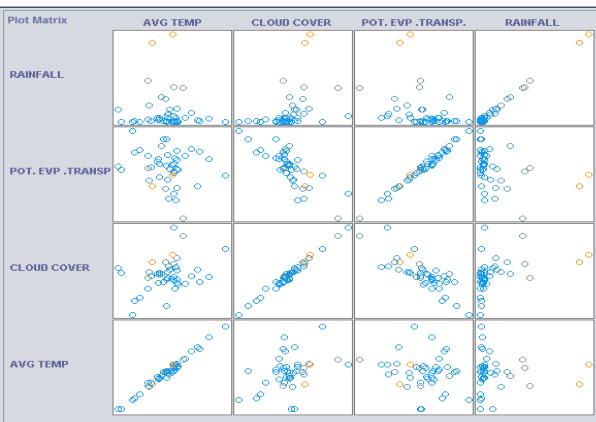
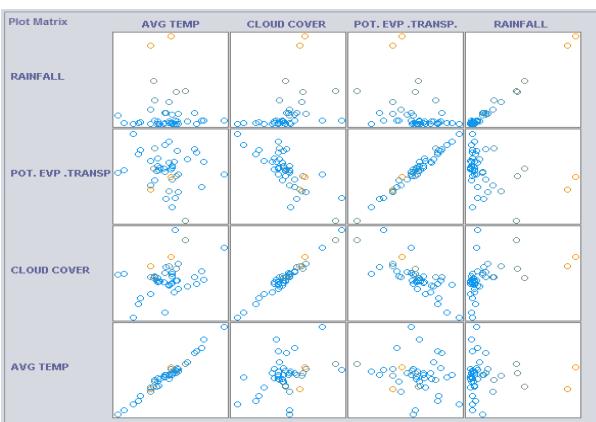


FIGURE 15. Result of SVM regression for November.

5) MONTH 5(MAY):NON-MONSOON REGRESSION ANALYSIS

The performance results of both SVM (RBF Kernel) model by 70% train and 30% test and MLP(2,2) model using 75% train and 25% test split for April rainfall prediction using multiple regression analysis is shown in figures below:

**FIGURE 16.** Result of MLP regression for November.**FIGURE 17.** Result of SVM regression for December.**FIGURE 18.** Result of SVM regression for December.

6) MONTH 6(OCTOBER):NON-MONSOON REGRESSION ANALYSIS

The performance results of both SVM (Poly Kernel) model and MLP (2, 2) model for April rainfall prediction using multiple regression analysis is shown in figures below.

TABLE 1. Annual comparative study of regression result using SVR and MLP.

MODEL	Corelation Coefficient	MAE	RMSE
SVM(PUK Kernel)	0.9989	0.085	0.1024
MLP	0.9913	0.169	0.1874

TABLE 2. Non-monsoon rainfall regression result for various SVM Kernels.

SL NO	Non-monsoon Month	SVM	R	MAE	RMSE
1	JANUARY	RBF KERNEL	0.69	0.05	0.09
2	FEBRUARY	PUK KERNEL	0.68	0.09	0.13
3	MARCH	POLY KERNEL	0.62	0.12	0.15
4	APRIL	POLY KERNEL	0.59	0.20	0.22
5	MAY	RBF KERNEL	0.76	0.10	0.13
6	OCTOBER	POLY KERNEL	0.80	0.13	0.17
7	NOVEMBER	POLY KERNEL	0.79	0.07	0.11
8	DECEMBER	RBF KERNEL	0.71	0.07	0.13

TABLE 3. Non-monsoon rainfall regression result for various MLP models.

SL NO	Non-monsoon Months	R	MAE	RMSE
1	JANUARY	0.78	0.0971	0.1355
2	FEBRUARY	0.79	0.0981	0.1368
3	MARCH	0.82	0.1077	0.1377
4	APRIL	0.40	0.2047	0.2550
5	MAY	0.75	0.0882	0.1061
6	OCTOBER	0.83	0.1811	0.2084
7	NOVEMBER	0.70	0.1213	0.1454
8	DECEMBER	0.83	0.0716	0.0909

The data set was split by 75% train and 25% test for both models.

7) MONTH 7(NOVEMBER):NON-MONSOON REGRESSION ANALYSIS

Result of performance of both SVM (Poly Kernel) model and MLP (2, 2) model for April rainfall prediction using multiple regression analysis is shown in figures below.

8) MONTH 8(DECEMBER):NON-MONSOON REGRESSION ANALYSIS

Regression analysis was performed using SVM, RBF KERNEL. The result snapshots and plots of prediction

TABLE 4. Non-monsoon regression comparative study between SVM and MLP.

SL NO	Non-monsoon month	Accurate Model For Rain Fall prediction
1	JANUARY	SVM RBF KERNEL
2	FEBRUARY	SVM PUK KERNEL
3	MARCH	MLP(2,2)
4	APRIL	SVM POLY KERNEL
5	MAY	MLP(2,2)
6	OCTOBER	SVM POLY KERNEL
7	NOVEMBER	SVM POY KERNEL
8	DECEMBER	MLP(3,3)

using regression for December using SVM. As are shown in figures below.

VIII. ANALYSIS AND DISCUSS

The tabular comparative result for annual as well as non-monsoon regression is shown below:

A. THE RESULT OF MULTIPLE REGRESSIONS FOR ANNUAL RAINFALL PREDICTION IN ODISHA

The result of multiple regressions for annual rainfall prediction in Odisha is shown in table 1:

B. THE RESULT OF MULTIPLE REGRESSIONS FOR NON-MONSOON RAINFALL PREDICTION IN ODISHA

The result of multiple regressions for non-monsoon rainfall prediction in Odisha is shown in table 2, 3 and 4:

IX. CONCLUSION

This study shows that for annual rainfall prediction for Odisha using multiple regression analysis SVM (PUK Kernel) model showed better performance than MLP. For non-monsoon rainfall prediction using multiple regression analysis for 5 non-monsoon months, January, February, April, October and November the rainfall prediction will be more accurate if SVR is used for future rainfall prediction. For rest 3 non-monsoon months, March, May and December the future rainfall prediction will be more accurate if Multi Layer Perceptron is used. In future multiple or multi variate time series prediction can be used for more clarity of prediction. LSTM (Long Short-Term memory) can also be implemented in future with a hope of getting more accurate predictions. A variety of other parameters can be implemented if they are available.

ACKNOWLEDGMENT

The authors are thankful to the Centre for Environmental Studies (CES), Department of Forest and Environment, Government of Odisha, and to the website <https://www.indiawaterportal.org/> for the data.

REFERENCES

- R. Samsudin, A. Shabri, and P. Saad, "A comparison of time series forecasting using support vector machine and artificial neural network model," *J. Appl. Sci.*, vol. 10, no. 11, pp. 950–958, Nov. 2010.
- S. K. Trivedi and S. Dey, "Effect of various kernels and feature selection methods on SVM performance for detecting email spams," *Int. J. Comput. Appl.*, vol. 66, no. 21, pp. 18–23, 2013.
- B. Üstün, W. Melssen, and L. Buydens, "Facilitating the application of support vector regression by using a universal Pearson VII function based kernel," *Chemometric Intell. Lab. Syst.*, vol. 81, no. 1, pp. 29–40, Mar. 2006.
- S. M. Paras, A. Kumar, and M. Chandra, "A feature based neural network model for weather forecasting," *Int. J. Comput. Intell.*, vol. 4, no. 3, pp. 209–216, 2009.
- T. Koskela, M. Lehtokangas, J. Saarinen, and K. Kaski, "Time series prediction with multilayer perceptron, FIR and Elman neural networks," in *Proc. World Congr. Neural Netw.* San Diego, CA, USA: INNS Press, 1996, pp. 491–496.
- A. D. Kumarasiri and D. U. J. Sonnadara, "Rainfall forecasting: An artificial neural network approach," in *Proc. Tech. Sessions*, Sri Lanka, 2006, pp. 1–16.
- S. Chattopadhyay and M. Chattopadhyay, "A soft computing technique in rainfall forecasting," 2007, *arXiv:nlin/0703042*. [Online]. Available: <https://arxiv.org/abs/nlin/0703042>
- N. Q. Hung, M. S. Babel, S. Weesakul, and N. K. Tripathi, "An artificial neural network model for rainfall forecasting in Bangkok, Thailand," *Hydrol. Earth Syst. Sci.*, vol. 13, no. 8, pp. 1413–1425, 2009.
- R. R. Deshpande, "On the rainfall time series prediction using multilayer perceptron artificial neural network," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 1, pp. 2250–2459, 2012.
- K. Abhishek, A. Kumar, R. Ranjan, and S. Kumar, "A rainfall prediction model using artificial neural network," in *Proc. IEEE Control Syst. Graduate Res. Colloq.*, Jul. 2012, pp. 82–87.
- H. Mislan, S. Hardwinarto, and M. A. Sumaryono, "Rainfall monthly prediction based on artificial neural network: A case study in Tenggarong Station, East Kalimantan-Indonesia," *Procedia Comput. Sci.*, vol. 59, pp. 142–151, Jan. 2015.
- A. Ddubey, "Artificial neural network models for rainfall prediction in pondicherry," *Int. J. Comput. Appl.*, vol. 120, no. 3, pp. 30–35, Jun. 2015.
- N. Sharma, M. Zakullah, H. Tiwari, and D. Kumar, "Runoff and sediment yield modeling using ANN and support vector machines: A case study from Nepal watershed," *Model. Earth Syst. Environ.*, vol. 1, no. 3, p. 23, 2015.
- S. M. Sumi, M. F. Zaman, and H. Hirose, "A rainfall forecasting method using machine learning models and its application to the Fukuoka city case," *Int. J. Appl. Math. Comput. Sci.*, vol. 22, no. 4, pp. 841–854, Dec. 2012.
- S. Lee, J.-C. Kim, H.-S. Jung, M. J. Lee, and S. Lee, "Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea," *Geomatics, Natural Hazards Risk*, vol. 8, no. 2, pp. 1185–1203, Dec. 2017.
- S. Das, R. Chakraborty, and A. Maitra, "A random forest algorithm for nowcasting of intense precipitation events," *Adv. Space Res.*, vol. 60, no. 6, pp. 1271–1282, Sep. 2017.
- Y. Zaman, "Machine learning model on rainfall—A predicted approach for Bangladesh," Ph.D. dissertation, United Int. Univ., Dhaka, Bangladesh, 2018.
- S. S. Monira, Z. M. Faisal, and H. Hirose, "Comparison of artificially intelligent methods in short term rainfall forecast," in *Proc. 13th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2010, pp. 39–44.
- P. Supriya, M. Krishnaveni, and M. Subbulakshmi, "Regression analysis of annual maximum daily rainfall and stream flow for flood forecasting in Vellar river basin," *Aquatic Procedia*, vol. 4, pp. 957–963, 2015.
- Q. Liu, Z. Yanyun, X. Liu, and N. Linge, "A survey on rainfall forecasting using artificial neural," *Int. J. Embedded Syst.*, vol. 11, no. 2, pp. 12–19, 2019.
- S. A. Soleymani, S. Goudarzi, M. H. Anisi, W. H. Hassan, M. Y. I. Idris, S. Shamshirband, N. M. Noor, and I. Ahmedy, "A novel method to water level prediction using RBF and FFA," *Water Resour. Manage.*, vol. 30, no. 9, pp. 3265–3283, Jul. 2016.
- J. Wu, J. Long, and M. Liu, "Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm," *Neurocomputing*, vol. 148, pp. 136–142, Jan. 2015.



XIAOBO ZHANG was born in Suizhou, China, in 1977. He received the master's degree in computer science and the Ph.D. degree in engineering (control theory and control engineering). He is currently working in the Internet of Things at the Department of Automation, Guangdong University of Technology. He is also a Guangdong Science and Technology Commissioner. He is engaged in teaching, technology research and development, and product design and scientific theoretical research in the field of industrial control and information technology services. His main research interests are hardware and software design for the IoT and big data, intelligent information processing and intelligent control (robot), and other industrial control systems. He has participated in more than 60 national and provincial research projects and corporate research projects, including the National Natural Science Foundations, the Natural Science Foundations of Guangdong Province, major science and technology projects in Guangdong Province, and a combination of production, education and research projects of the Ministry of Education of Guangdong Province. In recent years, he has published more than 30 articles in academic journals with important influence at home and abroad, including three books as a Chief Editor. He applied for more than 40 invention patents and obtained six software copyrights.



SACHI NANDAN MOHANTY received the Ph.D. degree from IIT Kharagpur, India, in 2014, with the MHRD Scholarship from the Government of India. He has recently joined as an Associate Professor at the Department of Computer Science and Engineering, ICFAI Foundation For Higher Education, Hyderabad. His research areas include data mining, big data analysis, cognitive science, fuzzy decision making, brain-computer interface, cognition, and computational intelligence. He has published 15 international journals of international repute and four patents. He has been elected as a member of the Institute of Engineers and the IEEE Computer Society. He has received three best paper awards during his Ph.D. at IIT Kharagpur from the International Conference, Beijing, China, and other awards at the International Conference on Soft Computing Applications organized by IIT Roorkee, in 2013. He received the Best Thesis Award First Prize by the Computer Society of India, in 2015. He is also a Reviewer of IJAP and IJDM International Journals.



AJAYA KUMAR PARIDA submitted his Ph.D. thesis at Utkal University, India, in 2019. He is currently working as an Associate Professor at the Department of Computer Science and Engineering, KIIT University, Odisha. His research areas include data mining, big data analysis, and computational intelligence. He has published ten international journals of international repute. He has been elected as a member of the Institute of Engineers and the IEEE Computer Society.



SUBHENDU KUMAR PANI received the Ph.D. degree from Utkal University, Odisha, India, in 2013. He is currently working as a Professor at the Department of Computer Science and Engineering, and a Research coordinator at the Orissa Engineering College (OEC), Bhubaneswar. He has more than 16 years of teaching and research experience. His research interests include data mining, big data analysis, web data analytics, fuzzy decision making, and computational intelligence. In addition to research, he has guided one Ph.D. student and 31 M.Tech. students. He has published 51 international journal articles (19 Scopus index). He has more than 100 international publications, five authored books, two edited books, and ten book chapters into his account. He is a Fellow of SSARSC and a Life Member of IE, ISTE, ISCA, OBA.OMS, SMIACST, SMUACEE, and CSI. He was a recipient of five researcher awards. His professional activities include roles as an Associate Editor, an Editorial Board Member and/or a Reviewer of various international journals. He also holds associate positions with the number of conference societies.



BIN DONG received the B.S. degree in information and computing science and the M.S. degree in computer science and technology from North China Electric Power University, Baoding, China, in 2003 and 2006, respectively, and the Ph.D. degree in optical engineering from Hebei University, Baoding, in 2012. His research interests include big data analysis and the artificial intelligent algorithm in computer vision.



XIAOCHUN CHENG received the bachelor's degree in computer engineering and the Ph.D. degree in computer science from Jilin University, in 1992 and 1996, respectively. He has been a Computer Science EU Project Coordinator at Middlesex University, since 2012. He is a member of the IEEE SMC Technical Committee on Enterprise Information Systems, the IEEE SMC Technical Committee on Computational Intelligence, the IEEE SMC Technical Committee on Cognitive Computing, the IEEE SMC Technical Committee on Intelligent Internet Systems, the IEEE Communications Society Communications and Information Security Technical Committee, BCS Information Security Specialist Group, BCS Cybercrime Forensics Specialist Group, and BCS Artificial Intelligence Specialist Group.