Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques

Conference Paper · December 2018

DOI: 10.1109/PDGC.2018.8745763

5 authors, including:

Urmay Shah
Nirma University
1 PUBLICATION **8** CITATIONS
SEE PROFILE

Sanjay Garg
Indrashil University
103 PUBLICATIONS **410** CITATIONS
SEE PROFILE

Neha Sisodiya
2 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques

Urmay Shah
Dept. of Computer Engg.
Nirma University,
Ahmedabad, India
15mcen25@nirmauni.ac.in

Sanjay Garg
Dept. of Computer Engg.
Nirma University,
Ahmedabad, India
sgarg@nirmauni.ac.in

NehaSisodiya
Dept. of Computer Engg.
Nirma University,
Ahmedabad, India
neha.sisodiya@nirmauni.ac.in

Nitant Dube
SAC-ISRO
Ahmedabad, India
nitant@sac.isro.gov.in

Shashikant Sharma
SAC-ISRO
Ahmedabad, India
sasharma@sac.isro.gov.in

***Abstract*—The paper is focused to provide the insights of climate to the clients from various businesses, e.g. agriculturists, researchers etc., to comprehend the significance of changes in climate and atmosphere parameters like precipitation, temperature, humidity etc. Precipitation estimate is one of the critical investigations in field of meteorological research. In order to predict precipitation, an endeavor is made to a couple of factual procedures and machine learning techniques to forecast and estimate meteorological parameters. For experimentation purpose daily observations were considered. The accuracy assessment of forecasting model experimentation is done using validation of results with ground truth. The experimentation demonstrates that for forecasting meteorological parameters ARIMA and Neural Network works best, and best classification accuracy in comparison to other machine learning algorithms for forecasting precipitation for next season was given by Random Forest model.**

Keywords—Precipitation, ARIMA, SVM, Decision Tree, Holt Winter, Machine Learning, Random Forest

I. INTRODUCTION

In India, where the majority of agribusiness is dependent on precipitation as its standard wellspring of water, the time and measure of precipitation hold high importance and can impact the entire economy of the nation. Climate plays a vital role in our everyday life. From the earliest starting point of the human development, we are occupied with thinking about climatic changes. Weather forecasting is one of the most challenging issues seen by the world, in a most recent couple of century in the field of science and technology. Prediction is the phenomena of knowing what may happen to a system in the near future. Present weather observations are obtained by ground-based instruments and from the satellite through remote sensing. As India's economy significantly depends on horticulture, precipitation plays an important part.

The monthly climatic changes using spatiotemporal mining is being analyzed and the variability in seasonal rainfall using the IMD data with many rain gauge station information is done by K. Chowdari in[5][1]. Cluster analysis technique is also performed using no. of rainy days and rainfall as the input variable. L. Ingsrisawang in[11] has done a comparative study for rainfall prediction using diff-erent machine learning techniques on the north-eastern part of Thailand. The paper shows that, how the feature selection can be used to find the correlation between other weather parameter and the rainfall, the paper also shows the same day, next day, and next 2-day classification using ANN, SVM, KNN. Thai meteorological department (TMD) data is used for experimental purpose. Attributes like temperature, humidity, pressure, wind, rain occurrence are used as input to the model. In [15] S.N Kohail has used daily historical data of the Gaza city and outlier analysis, prediction, classification, and clustering is done for temperature prediction. The paper shows the temperature prediction and classification for the Gaza city using many machine learning techniques, it also does outlier detection and clustering. Daily relative humidity, average temperature, wind speed with direction, time of highest speed and rainfall is used as an input parameter in the study. Onset monsoon for the Indian sub-continent is predicted based on features extracted from the satellite image using data mining methods. KNN with euclidean distance is used for sea surface temperature (SST), cloud top temperature (CTT), cloud density, water vapour attributes were used. It predicts the onset monsoon in advance 10-30 days is proposed in[13].

Rainfall classification using supervised learning in Quest (SLIQ), and decision tree method with different Gini index is performed in[18]. Dew point, temperature, pressure, humidity, wind speed were used as an input parameter. Petre in [17](2008) proposed an approach that uses decision tree method with CART algorithm using data from meteorological department Hong Kong. They have used year, month, average pressure, relative humidity, cloud quantity, precipitation, average temperature as an input parameter. The work is explained by S.-Y. Ji in[12] uses decision tree with CART and C4.5 algorithm with temperature, wind direction, wind speed, wind gust, outdoor

humidity, evaporation, solar radiation, dew point, cloud cover, air density, vapour pressure, pressure altitude as a parameter. The proposed method predicts rain and it is classified into three categories in hourly rain 0.0 to 0.5 mm as level 1, 0.5 to 2.0 mm as level 2, > 2.0mm as level 3.

A comparative study of data mining techniques is being done using historical weather data set. Analysis of different machine learning techniques for regression as well as for the classification, paper shows that KNN performs better for classification and Naive Bayes performs better for regression [2]. Forecasting monthly rainfall for the Assam region using multiple linear regression is performed with the help of 6 years data gathered from regional meteorological center Guwahati[6]. Using Nigerian Meteorological Agency data paper individually predict the min temp, max temp, evaporation, rainfall and radiation using ANN and decision tree, error in rainfall is very high compared to other parameters prediction error in [14]. In [19] the comparison of several machine learning algorithms like ANN, Multiplicative Additive Regression Spline (MARS), radial basis SVM is done to forecast average daily and monthly rainfall of the Fukuoka city Japan. Rainfall forecasting using neural network through the satellite image is attempted in [14] with parameters like relative humidity, pressure, temperature, precipitate water, wind speed. Daily rainfall prediction over Dhaka station in Bangladesh using markov chain model and logistic regression is performed with the help of no of rainy days, no of dry days and rainfall as a parameter in [8][16].

We have observed that most of the papers [5][2][19][14][9][3] claiming higher accuracy have classified rainfall into three or less than three categories or have estimated rainfall using machine learning techniques but have not done rainfall forecasting using machine learning techniques, few of them have used few meteorological parameters for the estimation of the rainfall. The papers which are forecasting rainfall have used the regression techniques and forecasting techniques have less accuracy. We have proposed a model to predict the rainfall using a fusion of forecasting and machine learning techniques. Prediction of rainfall depends on various other parameters along with temperature. Classifying the rainfall gives us the good classification accuracy but our ultimate goal is to predict the rainfall using the other forecasted parameters. In this study objective is not only to correctly classify rainfall but also correctly predict the rainfall using various forecasted parameters. Our work is focused on understanding the effects of different meteorological parameters in rainfall prediction along with an exploration of approaches which were used for forecasting rainfall, machine learning, and their limitation. The proposed model predicts the rainfall for the next season using machine learning and forecasting techniques. Our contribution to this problem is to analyze the accuracy of different machine learning and forecasting techniques to predicts precipitation for next season.

The rest of paper is organized as follows: Section II elaborates the methods used. The proposed architecture is explained in Section III. Section IV presents the source of data information, parameters used with their dates are mentioned along with comparison of the results of the different machine learning and forecasting methods. This section also includes the final classification accuracy of the rainfall. We have concluded our work in Section V with future scope.

II. Basic Preliminaries

A. Machine learning methods for Regression

Multiple Linear Regression: In multiple linear regression[7], multiple in-dependent parameters are taken as an input and based on the best-fitted line dependent continuous variable is predicted. The relation between them is derived by equation:

$$Y=a*X+b*Z+c$$

Where Y =Dependent Variable, a,b =Regression Parameters X, Z=Independent Variable, c=Intercept

Support Vector Regression: The support vector regression (SVR) [3] uses the same principles as the SVM for classification, with only a few minor differences. To minimize error, individualizing the hyper plane which maximizes the margin, keeping in mind that part of the error is tolerated in linear support vector regression.

Prediction of the rainfall using other independent parameters (temperature, humidity, pressure, wind speed etc.) is attempted in many studies showing the comparison of different machine learning techniques and claiming the higher ac-curacy with categorizing rainfall in two to three categories, but most of them have not attempted the forecasting of rainfall for next season using machine learning techniques. In Few papers forecasting of the rainfall as well as different weather parameters like temperature, relative humidity, number of rainy days etc. is attempted. The result shows forecasting rainfall individually gives less accurate result compared to other weather parameters.

As forecasting rainfall individually using forecasting techniques gives less accuracy and prediction of rainfall with the help of different weather parameter using machine learning techniques gives higher accuracy it is necessary to design the fusion model.

III. Proposed Architecture

In the first part of the proposed model retrieved weather data is cleaned and reordered, after that the rainfall data is categorized into different categories according to IMD guidelines. The data is partitioned into two parts 70% for training and 30% for testing. Four different machine learning techniques like a decision tree, random forest, KNN, SVM were applied on the partitioned data, the individual results were also analyzed and tuned.

In the second part of the proposed model, the correlation of the rainfall with minimum temperature, maximum temperature, relative humidity and wind speed were calculated. From the study, it is found that all four parameters have significant importance with the rainfall. All past years maximum temperature and minimum temperature

were retrieved except last year. Based on the past data six different forecasting methods (Holt winter method [10], ARIMA model [10], Simple Moving Average model [2], Neural Network method [10], Seasonal Naive method [10]) were applied and the best-fitted model output was taken into consideration. Relative humidity and wind speed were retrieved from minimum temperature and maximum temperature using linear regression and support vector regression as it is found that it gives better accuracy by this method compared to a direct forecast of the individual.

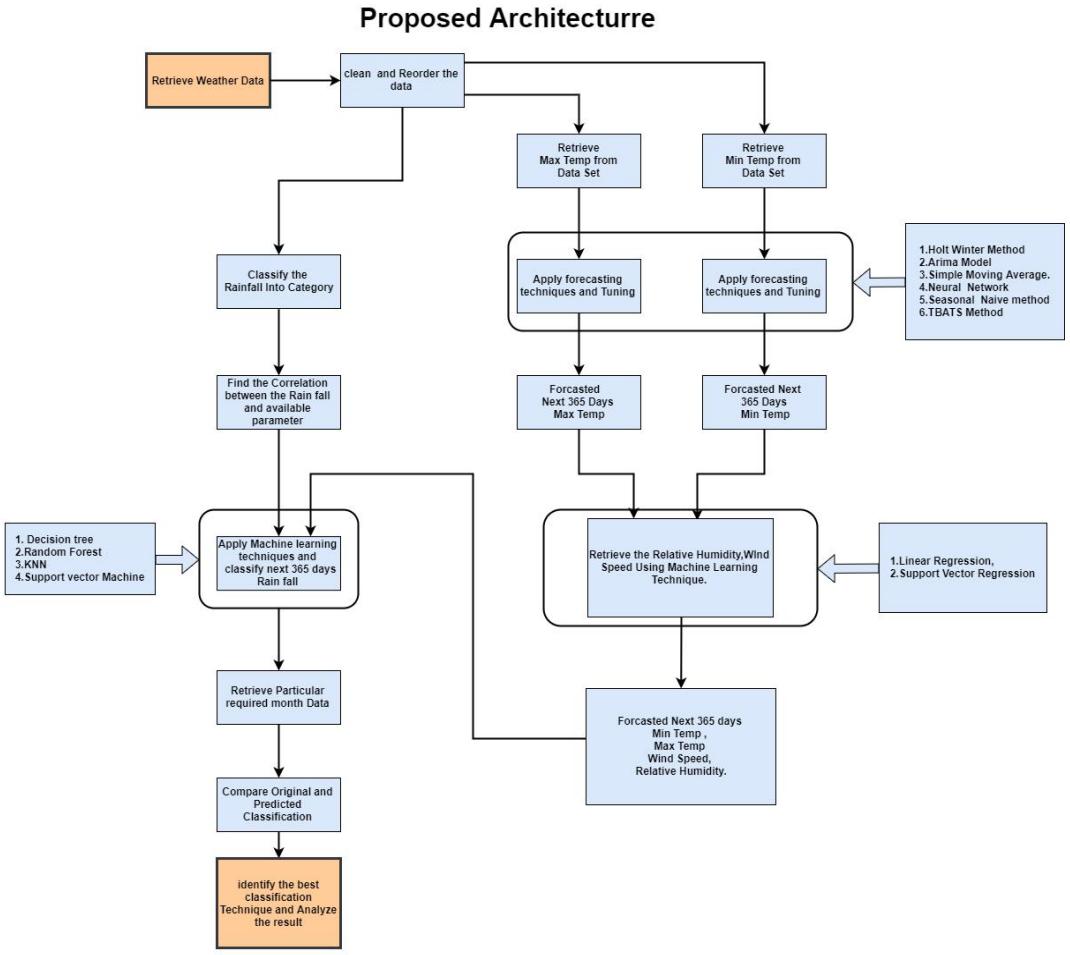


Fig. 1: Proposed Architecture

In the fusion part, four forecast parameters are given as input to the trained data (1979 to 2013). Based on this input parameters next year and next monsoon season rainfall is forecasted. The individual accuracy of the model was also analyzed with confusion matrix. For the experimental purpose we have taken only Jun to Dec data because in most of regions of India rainfall occurs in this period. Considering the forecast for whole year gives higher accuracy as there are more no. of non rainy days which gets correctly classified but our focus is to predict the rainfall for those months who have chances of rainfall.

IV. Results and Tables

This sections includes the information about the data which is used for the experimentation along with the results of the forecasting and machine learning methods with the detailed explanation of tuned parameters. The detailed analysis of the best-fitted model and comparison of all methods based on performance is done.

The data for the experimental purpose is retrieved from global weather site and it is provided by National Centers for Environmental Prediction (NCEP). For experimentation, daily data from 1/1/1979 to 7/31/2014 is collected from five locations. Data also contains parameters like minimum temperature, maximum temperature, relative humidity, wind speed, and precipitation. The rainfall is classified into seven categories according to the forecast manual provided by the Indian Meteorological Department (IMD).Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Forecasting Parameters

Forecasting Maximum Temperature: As the temperature is comparatively easy to forecast compare to other meteorological parameters. We have forecasted maximum temperature using different forecasting techniques and RMSE (Root Mean Square Error) were compared with the

original data set. Table 1 shows, the 365 days forecasted maximum temperature error. The result shows that in the case of maximum temperature ARIMA model performs better than the other model. Arima(3,0,4) is the best-fitted model.

Forecasting Minimum Temperature: Various forecasting method for fore-casting the minimum temperature were analyzed, neural network show significant low RMSE compared to the other model. NNR(30,1,16)[365] performs the best fit model. Average of 20 networks, each of which is a 31-16-1 network with 529 weights options were -linear output units. Estimated sigma2 =0.01786

Table 1: Forecasted Maximum, Minimum Temperature RMSE

Method	RMSE (C)
ARIMA	3.45
TBATS Model	3.53
Naive Method	4.33
Moving Average	6.93
Neural Network	8.66
Holt Winters Additive	17.47
Holt Winters Multiplicative	13.57

Method	RMSE (C)
ARIMA	3.05
TBATS Model	3.57
Naive Method	3.38
Moving Average	7.92
Neural Network	2.55
Holt-Winters Additive	7.19
Holt-Winters Multiplicative	7.41

Forecasting Relative Humidity: As the correlation between relative humidity and rainfall is significant 0.303. We have also forecasted relative humidity. We have used minimum temperature and maximum temperature as the input to the model and predicted relative humidity. Forecasted minimum and maximum temperature were given as the input instead of the measured temperature to get the final model accuracy. The result shows that support vector regression which is a combination of linear regression and support vector machine works best.

Forecasting Wind Speed: Wind speed is one of the important parameter for predicting the rainfall as its correlation with the rainfall is 0.49. It is also impor-tant to forecast the wind speed(m/s). We have also applied the two regression techniques for predicting the wind speed giving two input parameters minimum temperature and maximum temperature, as a result support vector regression gives less RMSE compare to simple linear regression.For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

Table 2: Forecasted Relative Humidity and Wind Speed RMSE

Forecasted Relative Humidity		Forecasted Wind Speed	
Method	RMSE(Fraction)	Method	RMSE(m/s)
Linear Regression	0.75	Linear Regression	0.1345
Support Vector Regression	0.68	Support Vector Regression	0.1116

B. Machine Learning Model

KNN Method: To identify the best k nearest neighbor, we have tried with different values of K. The study reveals that k=15 gives best classification accuracy for the 1-year forecast, and k=9 gives best classification accuracy for June to December month forecast. Confusion matrix shows that very heavy rain clas-si ed to none. Results also show the considerable accuracy for the no rain, very light rain, moderate rain. For the very heavy rain, heavy rain and rather heavy rain results were not impressive.

Decision Tree: In this method, we have used Gini index algorithm for the selection of the most homogeneous node. Higher the value of Gini higher the homogeneity and based on that decision tree is generated.

The process of pruning is also done in order to limit the level of the tree. To ensure that tree is not overfitted or underfitted we have also tuned tree. For level 5, it shows the best result, to avoid overfitting, we have taken only up to 5 level.10-fold cross validation is done on this data set for measuring the accuracy of the model.

Results were also analyzed by confusion matrix. It is found that unlike the KNN, this method has classified very heavy rain. But same as the case in KNN it only shows the considerable accuracy for the no rain, moderate rain and for very light rain.

Support Vector Machine: In order to give best classification accuracy different combination of kernels, gamma, C values were tried for the tuning purpose. Radial base function kernel, linear kernel, sigmoid kernel were given for kernel parameter, different gamma values and C values were also given. It is found that linear kernel with gamma value 0.1 and C value 1 gives best accuracy compared to others. From the confusion matrix, it is found that SVM is unable to classify Heavy Rain and Very Heavy Rain. For even light rain and for very light rain results were poor.

In the experimentation we have taken more number of classes to classify the rainfall, but as SVM works best with optimal margin, there may be the case that multiple category overlap each other and because of which SVM performs worst compare to others.

Random Forest: Random forest [4] is a tree based model, it is a collection of many tree models. We have applied different tuning parameters for tuning it. As in random forest case, one of the parameters is how many trees should be used to get the more accurate results. It works well with high variance low bias models. It is noticed that after 250 number trees error rate is constant. So, we will restrict number of trees to 250 in the forest. From the confusion matrix, it is found that for very light rain Random forest method gives the best accuracy. It also performs well for the no rain, moderate rain, and for light rain.

Table 3: Accuracy on 30% test data.

Method	AUC (Area Under Curve)	Classification Accuracy	Precision	Recall
KNN	0.873	0.721	0.691	0.721
Tree	0.755	0.721	0.716	0.721
SVM	0.684	0.539	0.659	0.539
Random Forest	0.914	0.762	0.744	0.76

Table 4: Final Accuracy Comparison on Forecast

Method	Final Accuracy(1 year-365 days)	Final Accuracy (for June To Dec)
Decision Tree	69.58	61.21
Random Forest	70.50	70.09
KNN	69.31(k=15)	66.35(k=9)
SVM	67.05	69.15
Neural Network	68.49	68.69

Random forest uses their own sample of training data i.e. there are some observations which might appear several

times in the sample. The final prediction is based on voting by each tree in the forest. Random Forest is characterized by their efficiency to deal with large data set, relatively robustness for outliers and noise and ability to deal with highly correlated predictor variables.

C. Accuracy Measurements and Analysis

Table 3 is the result of 70% training and 30% training data set. It shows that random forest out performs compared to another method. For the experimental purpose, we have given actual Real time values of Maximum Temperature, Min temperature, Relative Humidity, Wind Speed as an input to the trained model and analysis is done on 30% testing data set. But as we want to forecast the rainfall it is necessary to give forecasted maximum temperature, minimum temperature, relative humidity and wind speed values as an input parameter to the trained model. This forecasted parameters also have their own error so if we put the forecasted value as input parameter to this classification technique there are chances to decrease the final accuracy of the model. As the random forest gives the best accuracy we have shown the final con-fusion matrix for the random forest only (for Jun to Dec). Table 5 shows the confusion matrix for the random forest. Diagonal shows the correctly classified category. It is found that it shows good accuracy for No rain, moderate rain, very light rain, light rain. Figure 2 shows the ROC (Receiver Operator Characteristics) Curve analyzed through comparing results of the different methods, category wise. Note: Individual graphs are drawn for True Positive Rate (Sensitivity) on y-axis against False Positive Rate (1-Speci city) on x-axis for each categories.

Table 5 shows the final classification accuracy of each method with forecasted parameters as an input to the trained model. In a country like India, where rainfall occurs in only limited no. of the month. So for that, we have also analyzed our accuracy for monsoon season and it is noticed that it gives considerable classification accuracy.

Table 5: Confusion Matrix for the Random Forest (for Jun to Dec)

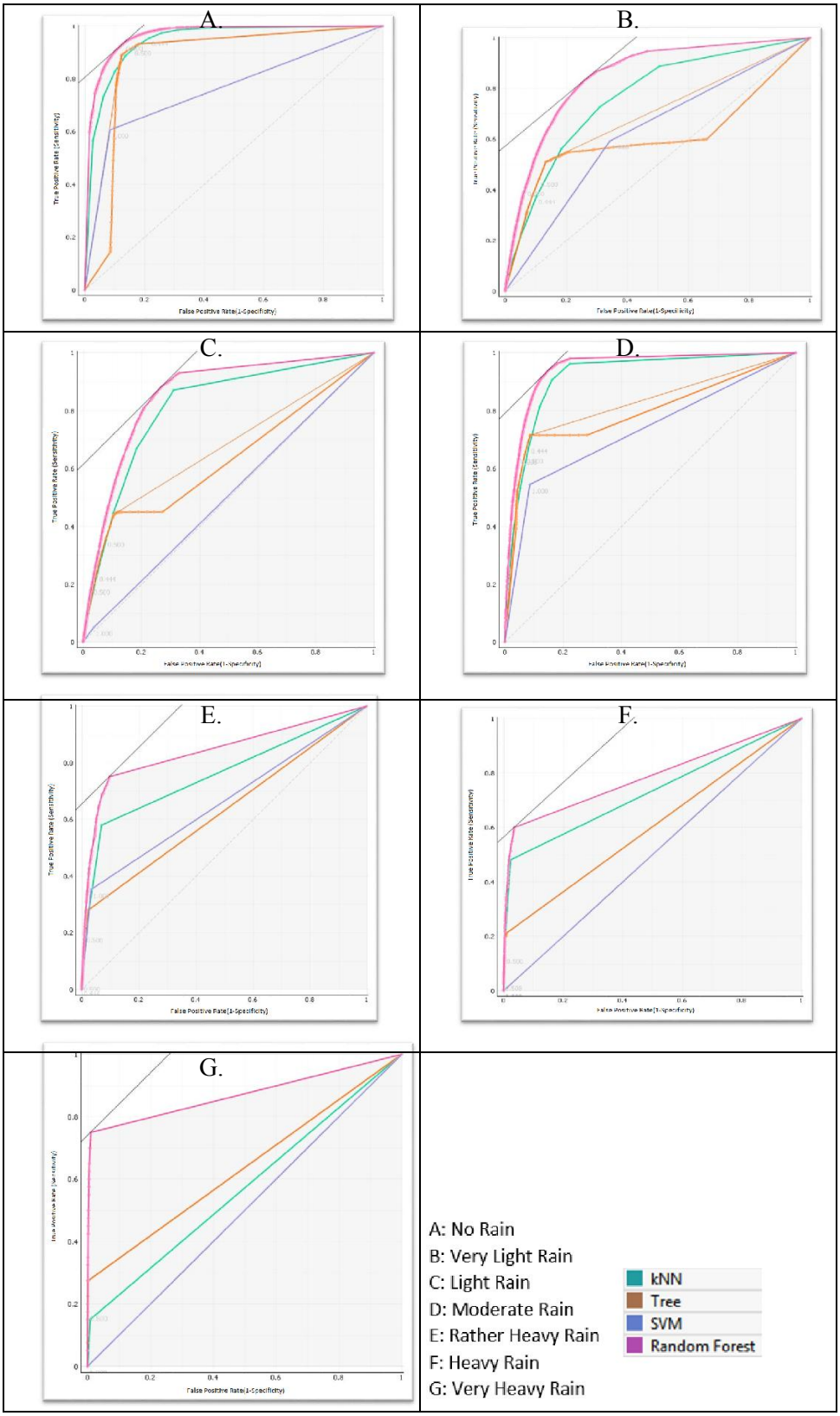
	Actual(Days)						
predicted (Days)	Heavy Rain	Light Rain	Moderate Rain	No Rain	Rather Heavy	Very Heavy Rain	Very Light Rain
Heavy Rain	0	0	0	0	0	0	0
Light Rain	0	12	4	0	1	0	8
Moderate Rain	0		33	0	1	1	9
No Rain	0	0	0	74	0	0	5
Rather Heavy	1	0	0	0	1	0	0
Very Heavy Rain	0	0	0	0	0	0	0
Very Light Rain	0	11	2	12	0	0	30

V CONCLUSION AND FUTURE WORK

The proposed work is an attempt to forecast rainfall using a fusion of different machine learning and forecasting techniques. Even though the rainfall is dependent on many parameters, we are able to get impressive classification accuracy using limited parameters. It is also found that even after classifying rainfall into eight different categories, we are getting acceptable accuracy. Validations for forecasted parameters are done using RMSE measure. Empirical results show ARIMA for maximum temperature, Neural Network for minimum temperature and SVR for relative humidity and wind speed works best. Validation of classification is measured through accuracy, precision and recall. ROC curve for all classifiers shows random forest works best for rainfall classification.

As rainfall is dependent on the various parameters it is also required to study how other meteorological parameters affect the Rainfall prediction. We can also perform the same exercise on hourly data using various parameters to forecast next hour rainfall. A study can also be done using more observations for particular region or area, and design this kind of model on big data framework so that computation can be faster with higher accuracy.

Fig. 2: ROC Curve Analysis Category wise



ACKNOWLEDGMENT

This work is financially supported by Space Application Center-Indian Space Research Organization (SAC-ISRO) under RESPOND Scheme OGP142. Authors would like to extend their sincere thanks to SAC-ISRO authorities for the opportunity given to work with them.

REFERENCES

[1] Mithila Sompura Aakash Parmar, Kinjal Mistree. Machine learning techniques for rainfall prediction: A review. International Conference on Innovations in informa-tion Embedded and Communication Systems, 2017.

[2] Nishchala C Barde and Mrunalinee Patole. Classification and forecasting of weather using ann, k-nn and na•ve bayes algorithms.

[3] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. Neural Information Processing-Letters and Reviews, 11(10):203{224, 2007.

[4] Leo Breiman. Random forests. Machine learning, 45(1):5{32, 2001.

[5] KK Chowdari, R Girisha, and KC Gouda. A study of rainfall over india using data mining. In Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on, pages 44{47. IEEE, 2015.

[6] Pinky Saikia Dutta and Hitesh Tahbilder. Prediction of rainfall using data mining technique over assam. IJCSE, 5(2):85{90, 2014.

[7] G Gregoire. Multiple linear regression. European Astronomical Society Publications Series, 66:45{72, 2014.

[8] Mina Mahbub Hossain and Sayedul Anam. Identifying the dependency pattern of daily rainfall of dhaka station in bangladesh using markov chain and logistic regression model. 2012.

[9] Rob J Hyndman. Moving averages. In International Encyclopedia of Statistical Science, pages 866{869. Springer, 2011.

[10] Rob J Hyndman and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2014.

[11] Lily Ingsrisawang, Supawadee Ingsriswang, Saisuda Somchit, Prasert Aung-suratana, and Warawut Khantiyanan. Machine learning techniques for short-term rain forecasting system in the northeastern part of thailand. Machine Learning, 887:5358, 2008.

[12] Soo-Yeon Ji, Sharad Sharma, Byunggu Yu, and Dong Hyun Jeong. Designing a rule-based hourly rainfall prediction model. In Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on, pages 303{308. IEEE, 2012.

[13] Dinu John and BB Meshram. A data mining approach for monsoon prediction using satellite image data. International Journal of Computer Science & Communication Networks, 2(3), 2012.

[14] Jyothis Joseph and TK Ratheesh. Rainfall prediction using data mining techniques. International Journal of Computer Applications, 83(8), 2013.

[15] Sarah N Kohail and Alaa M El-Halees. Implementation of data mining techniques for meteorological data analysis. Intl. Journal of Information and Communication Technology Research (JICT), 1(3), 2011.

[16] Folorunsho Olaiya and Adesesan Barnabas Adeyemo. Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1):51, 2012.

[17] Elia Georgiana Petre. A decision tree for weather prediction. BULETINUL UniversitaNii Petrol{Gaze din Ploiesti, pages 77{82, 2009.

[18] Narasimha Prasad, Prudhvi Kumar, and Naidu Mm. An approach to prediction of precipitation using gini index in sliq decision tree. In Intelligent Systems Modelling & Simulation (ISMS), 2013 4th International Conference on, pages 56{60. IEEE, 2013.

[19] Sirajum Monira Sumi, MFaisal Zaman, and Hideo Hirose. A rainfall forecasting method using machine learning models and its application to the fukuoka city case.