International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Rainfall Prediction Using Machine Learning

P. Suguna Reddy[1], *C. Sai Deepika*[2], *G. Pravalika*[3], *J. Jyoshna Devi*[4], *B.Kavya*[5]

[1]Department of Computer Science&Engineering, Santhiram Engineering College, India,
[2, 3, 4, 5, 6]Department of Computer Science&Engineering, Santhiram Engineering College, India

ABSTRACT—

The paper is targeted to provide the insights of weather to the clients from various companies, e.G. Agriculturists, researchers etc., to understand the importance of modifications in climate and environment parameters like precipitation, temperature, humidity etc. Precipitation estimate is one of the critical investigations in subject of meteorological research. So that you can predict precipitation, an enterprise is made to more than one genuine strategies and system learning techniques to forecast and estimate meteorological parameters. For experimentation motive daily observations have been considered. The accuracy evaluation of forecasting version experimentation is achieved using validation of outcomes with ground reality. The experimentation demonstrates that for forecasting meteorological parameters ARIMA and Neural community works nice, and great class accuracy in contrast to different gadget gaining knowledge of algorithms for forecasting precipitation for next season become given by using Random wooded area version.

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

I. INTRODUCTION

In India, in which the majority of agribusiness is depending on precipitation as its preferred wellspring of water, the time and degree of precipitation maintain excessive significance and may effect the whole financial system of the state. Climate performs a crucial function in our regular existence. From the earliest starting point of the human improvement, we are inquisitive about considering climatic changes. Climate forecasting is one of the maximum challenging troubles visible by means of the arena, in a maximum recent couple of century within the area of technological know-how and generation. Prediction is the phenomena of understanding what may also appear to a device inside the near destiny. Gift climate observations are obtained with the aid of ground-primarily based devices and from the satellite through far off sensing. As India's financial system extensively relies upon on horticulture, precipitation plays an vital part.

The month-to-month climatic modifications using spatiotemporal mining is being analyzed and the variety in seasonal rainfall the use of the IMD records with many rain gauge station information is achieved by using k. Chowdari in[5][1]. Cluster

Cluster analysis method is also done the usage of no. Of wet days and rainfall as the enter variable. L. Ingsrisawang in[11] has achieved a comparative examine for rainfall prediction the use of diff- erent machine gaining knowledge of techniques at the north-jap part of Thailand. The paper shows that, how the characteristic selection may be used to locate the correlation among different weather parameter and the rainfall, the paper additionally indicates the same day, next day, and next 2-day category the use of ANN, SVM, KNN. Thai meteorological branch (TMD) facts is used for experimental cause. Attributes like temperature, humidity, stress, wind, rain occurrence are used as input to the model. In [15] S.N Kohail has used each day historic records of the Gaza metropolis and outlier evaluation, prediction, type, and clustering is finished for temperature prediction. The paper indicates the temperature prediction and class for the Gaza metropolis the use of many device getting to know strategies, it additionally does outlier detection and clustering. Day by day relative humidity, common temperature, wind speed with course, time of maximum pace and rainfall is used as an input parameter inside the examine. Onset monsoon for the Indian sub-continent is expected based totally on features extracted from the satellite image using statistics mining techniques. KNN with euclidean distance is used for sea floor temperature (SST), cloud top temperature (CTT), cloud density, water vapour attributes had been used. It predicts the onset monsoon in advance 10-30 days is proposed in[13].

Rainfall classification the usage of supervised getting to know in Quest (SLIQ), and selection tree method with exceptional Gini index is done in[18]. Dew point, temperature, pressure, humidity, wind pace have been used as an enter parameter. Petre in [17](2008) proposed an technique that makes use of choice tree approach with CART algorithm the usage of statistics from meteorological branch Hong Kong. They have used yr, month, common stress, relative humidity, cloud amount, precipitation, average temperature as an input parameter. The paintings is defined by way of S.-Y. Ji in[12] makes use of decision tree with CART and C4.Five algorithm with temperature, wind path, wind pace, wind gust, Out of doors humidity, evaporation, sun radiation, dew factor, cloud cover, air density, vapour pressure, stress altitude as a parameter. The proposed method predicts rain and it's miles classified into 3 classes in hourly rain 0.Zero to 0.Five mm as degree 1, 0.5 to two.Zero mm as degree 2, > 2.0mm as stage 3.

A comparative observe of statistics mining techniques is being executed using ancient climate statistics set. Evaluation of different device learning strategies for regression in addition to for the class, paper indicates that KNN performs better for type and Naive Bayes plays higher for regression [2]. Forecasting monthly rainfall for the Assam region the usage of a couple of linear regression is completed with the assist of 6 years records accrued from regional meteorological middle Guwahati[6]. The use of Nigerian Meteorological agency statistics paper in my opinion expect the min temp, max temp, evaporation, rainfall and radiation the use of ANN and selection tree, error in rainfall may be very high compared to different parameters prediction error in [14]. In [19] the comparison of several system gaining knowledge of algorithms like ANN, Multiplicative Additive Regression Spline (MARS), radial foundation SVM is performed to forecast average day by day and month-to-month rainfall of the Fukuoka town Japan. Rainfall forecasting the use of neural community via the satellite tv for pc photo is tried in [14] with parameters like relative humidity, pressure, temperature, precipitate water, wind velocity. Day by day rainfall prediction over Dhaka station in Bangladesh using markov chain model and logistic regression is achieved with the help of no of wet days, no of dry days and rainfall as a parameter in [8][16].

We've located that most of the papers [5][2][19][14][9][3] claiming higher accuracy have categorized rainfall into 3 or much less than three classes or have expected rainfall the usage of machine learning strategies but have now not carried out rainfall forecasting the usage of system studying techniques, few of them have used few meteorological parameters for the estimation of the rainfall. The papers which might be forecasting rainfall have used the regression techniques and forecasting strategies have less accuracy. We've got proposed a model to are expecting the rainfall thae usage of a fusion of forecasting and machine getting to know techniques. Prediction of rainfall depends on numerous different parameters together with temperature. Classifying the rainfall gives us the good class accuracy however our remaining goal is to predict the rainfall the usage of the other forecasted parameters. On this look at objective is not only to correctly classify rainfall but additionally efficiently expect the rainfall the use of various forecasted parameters. Our work is targeted on know-how the outcomes of different meteorological parameters in rainfall prediction together with an exploration of procedures which had been used for forecasting rainfall, device getting to know, and their obstacle. The proposed version predicts the rainfall for the following season the use of machine learning and forecasting strategies. Our contribution to this problem is to analyze the accuracy of various device mastering and forecasting techniques to predicts precipitation for next season.

The relaxation of paper is prepared 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

A couple of Linear Regression: In a couple of linear regression[7], more than one in-structured parameters are taken as an input and primarily based at the exceptional-equipped line established continuous variable is anticipated. The relation between them is derived by equation:

Y=a*X+b*Z+c

In which Y =dependent Variable, a,b =Regression Parameters X, Z=impartial Variable, c=Intercept

Guide Vector Regression: The guide vector regression (SVR) [3] uses the equal standards as the SVM for classification, with only some minor variations. To decrease error, individualizing the hyper plane which maximizes the margin, maintaining in thoughts that part of the mistake is tolerated in linear aid vector regression.Prediction of the rainfall using other unbiased parameters (temperature, humidity, pressure, wind speed and so on.) is tried in lots of research showing the evaluation of different system mastering strategies and claiming the better ac-curacy with categorizing rainfall in to three categories, but maximum of them have no longer attempted the forecasting of rainfall for subsequent season the usage of machine mastering techniques. In Few papers forecasting of the rainfall in addition to different climate parameters like temperature, relative humidity, quantity of rainy days and so forth. Is tried. The result shows forecasting rainfall in my view offers much less correct result in comparison to other climate parameters.

As forecasting rainfall for my part the usage of forecasting techniques gives a great deal much less accuracy and prediction of rainfall with the assist of numerous weather parameter the use of system mastering strategies offers higher accuracy it's far important to layout the fusion model.PROPOSED architecture

In the first part of the proposed model retrieved weather statistics is wiped clean and reordered, after that the rainfall information is classified into one of a kind categories according to IMD guidelines. The information is partitioned into two parts 70% for training and 30% for testing. 4 exclusive device gaining knowledge of strategies like a choice tree, random forest, KNN, SVM had been implemented at the partitioned information, the person outcomes had been also analyzed and tuned.

Within the 2d a part of the proposed version, the correlation of the rainfall with minimum temperature, maximum temperature, relative humidity and wind pace have been calculated. From the examine, it's miles located that every one four parameters have massive importance with the rainfall. All past years maximum temperature and minimum temperature Were retrieved except remaining 12 months. Based on the beyond records six extraordinary forecasting techniques (Holt iciness technique [10], ARIMA model [10], simple moving average version [2], Neural network approach [10], Seasonal Naive technique [10]) had been applied and the first-rate-equipped version output turned into taken into consideration. Relative humidity and wind speed have been retrieved from minimal temperature and maximum temperature the use of linear regression and support vector regression as it's far determined that it gives better accuracy via this technique compared to an immediate forecast of the individual.

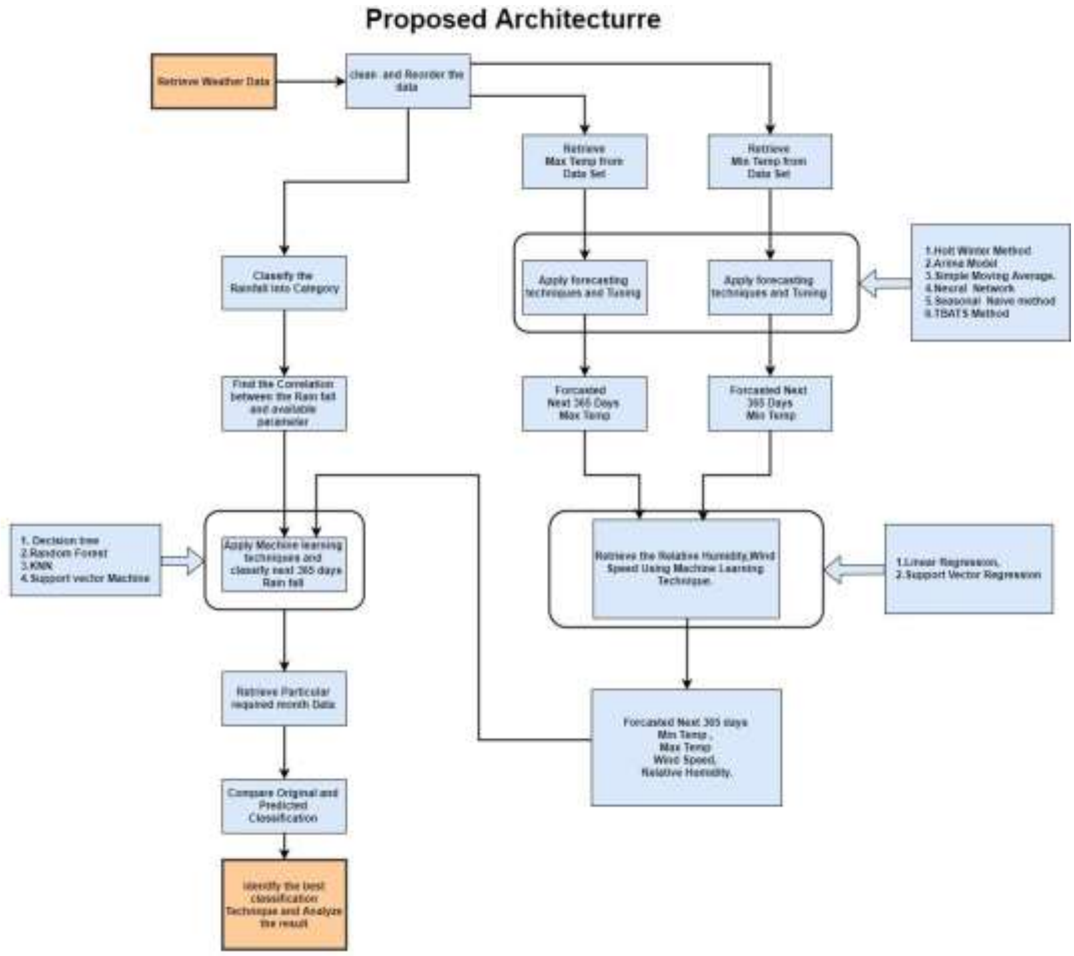

Proposed Architecturre
Retrieve Weather Data
clean and Reorder the data
Retrieve Max Temp from Data Set
Retrieve Min Temp from Data Set
Classify the Rainfall into Category
Apply forecasting techniques and Tuning
Apply forecasting techniques and Tuning
1.Holt Winter Method
2.Arima Model
3.Simple Moving Average.
4.Neural Network
5.Seasonal Naive method
6.TBATS Method
Find the Correlation between the Rain fall and available parameter
Forcasted Next 365 Days Max Temp
Forcasted Next 365 Days Min Temp
1. Decision tree
2.Random Forest
3.KNN
4.Support vector Machine
Apply Machine learning techniques and classify next 365 days Rain fall
Retrieve the Relative Humidity,Wind Speed Using Machine Learning Technique.
1.Linear Regression,
2.Support Vector Regression
Retrieve Particular required month Data
Forcasted Next 365 days Min Temp , Max Temp Wind Speed, Relative Humidity.
Compare Original and Predicted Classification
Identify the best classification Technique and Analyze the result

Fig. 1: Proposed Architecture

Within the fusion element, 4 forecast parameters are given as enter to the educated records (1979 to 2013). Based totally on this input parameters next 12 months and next monsoon season rainfall is forecasted. The man or woman accuracy of the model turned into also analyzed with confusion matrix. For the experimental purpose we have taken only Jun to Dec information because in most of regions of India rainfall occurs in this period. Considering the forecast for complete yr gives higher accuracy as there are more no. Of non rainy days which gets efficaciously labeled however our focus is to expect the rainfall for the ones months who've possibilities of rainfall.

III. 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, most temperature, relative humidity, wind velocity, and precipitation. The rainfall is assessed into seven classes in line with the forecast manual furnished via the Indian Meteorological branch (IMD).Shop As command, and use the naming conference prescribed by way of your convention for the name of your paper. On this newly created record, spotlight all of the contents and import your prepared textual content report. You are now geared up to style your paper; use the scroll down window on the left of the MS phrase Formatting toolbar.*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

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

A. 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.

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

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 l
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.

B. Accuracy Measurements and Analysis

Table 3 is the end result of 70% education and 30% training records set. It suggests that random wooded area out performs as compared to any other technique. For the experimental cause, we have given actual actual time values of most Temperature, Min temperature, Relative Humidity, Wind pace as an enter to the trained model and evaluation is achieved on 30% checking out statistics set. However as we want to forecast the rainfall it is essential to give forecasted maximum temperature, minimum temperature, relative humidity and wind pace values as an input parameter to the skilled version. This forecasted parameters additionally have their personal error so if we put the forecasted price as enter parameter to this type technique there are possibilities to decrease the final accuracy of the model. Because the random woodland offers the excellent accuracy we've got proven the final con-fusion matrix for the random wooded area only (for Jun to Dec). Table 5 shows the confusion matrix for the random woodland. Diagonal shows the efficaciously labeled category. It's miles found that it shows suitable accuracy for No rain, moderate rain, very light rain, light rain. Figure 2 indicates the ROC (Receiver Operator traits) Curve analyzed via evaluating outcomes of the exceptional methods, class sensible.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)						
Pedicted (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

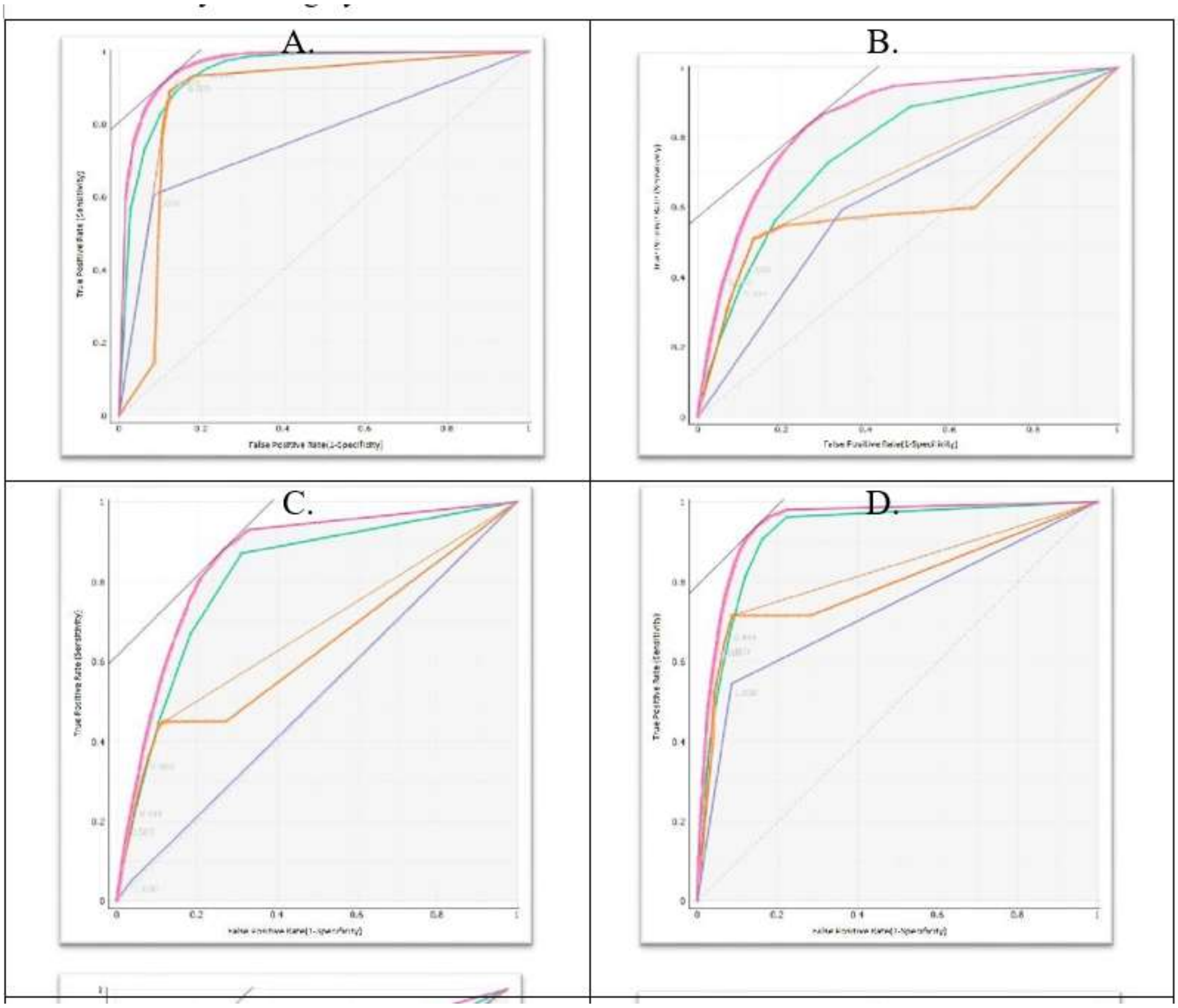
Very Heavy Rain	0	0	0	0	0	0	0
Very Light Rain	0	11	2	12	0	0	30

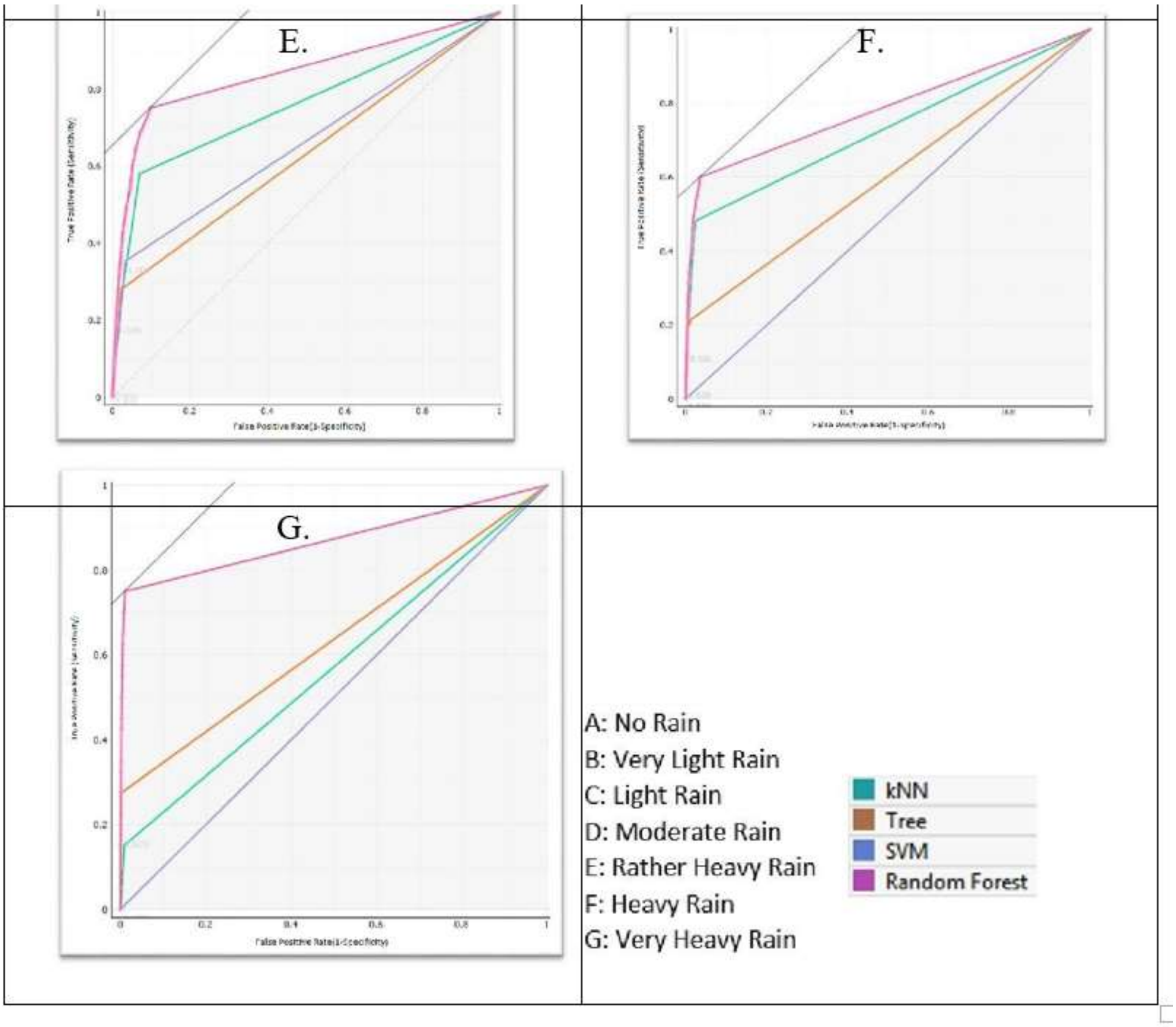
CONCLUSION AND FUTURE WORK

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 depending on the various parameters it's also required to look at how other meteorological parameters have an effect on the Rainfall prediction. We also can perform the identical exercising on hourly facts the usage of diverse parameters to forecast subsequent hour rainfall. A study can also be performed the use of more observations for particular region or region, and layout this sort of version on big statistics framework in order that computation may be faster with higher accuracy.

Fig. 2: ROC Curve Analysis Category wise





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 andforecasting 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.