

# Voting Classification Method for Rainfall Prediction

Maneesh Kumar  
Research Scholar

maneesh.dhmn@gmail.com,  
Computer Science Department  
Himachal Pradesh University Shimla

Dr. Jawahar Thakur  
Professor

jawahar.hpu@gmail.com  
Computer Science Department  
Himachal Pradesh University Shimla

**Abstract** - The rainfall is a significant part of water resource ecosystem and acts efficiently in the field of hydrology and meteorology. In particular, rainfall is the outcomes of multi-scale air system interference and various natural factors such as thermal power, flow field, and terrain have influence on it. The task to predict the rainfall becomes complex due to these complex physical mechanisms. In the process of forecasting rainfall, the probability of precipitation present in a certain region is predicted and rainfall in future is foreseen along with the estimation of the amount of rainfall in particular regions. The rainfall prediction has various phases which include pre-processing, feature extraction and classification. In this research work, voting classification method is designed for the rainfall prediction.

**Keywords** - Rainfall, Feature Selection, Voting Classification

## I. INTRODUCTION

The rainfall is a significant part of water resource ecosystem and acts efficiently in the field of hydrology and meteorology. In particular, rainfall is the outcomes of multi-scale air system interference and various natural factors such as thermal power, flow field, and terrain have influence on it. The task to predict the rainfall becomes complex due to these complex physical mechanisms. In the process of forecasting rainfall, the probability of precipitation present in a certain region is

predicted and rainfall in future is foreseen along with the estimation of the amount of rainfall in particular regions. The accuracy and error of prediction is considered in this process and the rainfall volume is estimated along with probability of rainfall in particular region. For this, the predictors collect, analyze, model [1], simulate and conduct research on diverse meteorological data and metrics. Some basic metrics such as average monthly temperature, amount of rain in and relative humidity. In addition, to predict the rainfall is related to daily life. The estimation of probability of rainfall along with its intensity assists the farmers and people in being aware of any disaster that leads to harm life and property. Data mining techniques have brought a major change in the traditional way of weather prediction. In the last few years, researchers have developed and practiced many weather forecasting models using data mining methodologies. These forecasting models have shown great accuracy in weather prediction [2]. Precipitation prediction with data mining technology, which is different from classic techniques of weather forecasting, has drawn a lot of attention from the research community. Based on machine learning theory, historical observational data can be used to predict future rainfall. Contrary to other types of models, this computing operation is clearly more suitable. There have been many worthwhile studies applying historical data to predict rainfall. Figure 1 illustrates a generic framework of rainfall prediction [3].

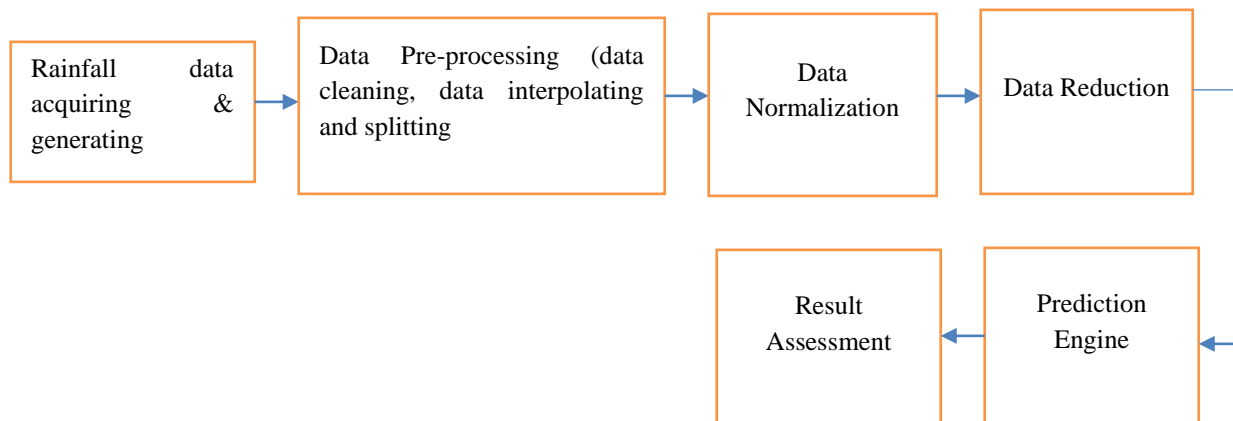


Fig. 1. Generic Framework of the Rainfall Prediction

All steps involved in the designing of a generic rainfall prediction framework have been described below:

i. Data Acquisition: This phase is focused on gathering the primary data for this study across India. Diverse metrics such as Average Monthly Temperature, Relative Humidity and annual rainfall are some metrics which are contained in the data. These metrics are taken from CES (Centre for Environmental Studies), from the Department of Forest and Environment, Government of India. The non-monsoon regression is analyzed monthly on the basis of weather metrics such as average temperature, cloud cover, potential evapo transpiration and rainfall [4].

ii. Pre-processing: The initial stage to pre-process the data is executed to attain the rainfall data and to execute the generating procedure. An automatic tool is constructed with the objective of obtaining and generating the data from the online data source. When the data is achieved and produced, the row data is cleaned, interpolated and divided. The raw datasets are consisted of some empty items and some duplicates. These duplicated and empty items are detected [5]. Various empty values transmission around the whole dataset is handled in the data interpolating stage. The fundamental goal of this phase is to normalize the min-max and mitigate the data. The average value is calculated from its chronicle neighbours and inserted as the estimated value. As an essential step for training and testing in machine learning models, data splitting is applied to divide the dataset into an appropriate proportion (generally in 70:30 ratio)

iii. Normalization: Z-score and Minmax schemes are adopted for computing the effect of normalization and non-normalization in the framework. The computation of Z-score value denoted with  $z_i$  is done as:

$$z_i = \frac{x_i - \mu}{\delta}$$

In which,  $x_i$  represents  $i - th$  observed value, the mean of all values is illustrated by  $\mu$  in the variable and  $\delta$  denotes the standard deviation in the variable [6]. Min-max is an effective technique utilized to normalize the data. This technique standardized the data amid 0 and 1. The data must be pre-processed as diverse factors contain distinct magnitude. To process a sequence, the maximum value of the sequence is corresponding to 1, the minimum value has to be 0, and the rest values are transformed proportionally amid 0 and 1. The min-max normalization can be mathematically expressed as:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In this,  $x$  is utilized to denote a value in the sequence of primitive variables,  $x_{max}$  illustrates the maximum values and  $x_{min}$  represents the minimum values in variables.

iv. Dimensionality reduction: After normalization, dimensionality of the input data is reduced. PCA is a common theory executed to pre-process the data and mitigate the dimension and extensively utilized in multivariate statistics.

The processing becomes more complicated with the maximization of number of variables. Principal Component Analysis is capable of displaying more information with fewer variables [7]. In case, the correlation coefficient among the initial variables is not 0, this implies a certain overlap among these variables. For all primary variables, PCA is effective for deleting the frequent relations of variables and generating novel variables in least amount. These new variables are independent of each other. However, the original information is stored. The vector space composed of input samples is converted to create the direction of the largest error as the base vector of the new linear space. Actually, PCA has potential to provide more information with relatively fewer factors. The complexity of the problem is simplified and the usage of hardware resources is alleviated using this algorithm[8]. The computation formula of this algorithm is:

$$Z_i = \mu_{i1}X_1 + \mu_{i2}X_2 + \dots + \mu_{ip}X_p, 1 \leq i \leq p$$

In this,  $p$  is the dimensional random vector,  $X = (X_1, X_2, \dots, X_p)_t$  is the original variable. The  $t$  moment is defined with  $t$ . A LT (linear transformation) is executed for  $X$ , and the conversion of original variable  $X$  is done into new variables  $Z = (Z_1, Z_2, \dots, Z_p)_t$ . A suitable coefficient  $\mu$  is employed to make it impossible to correlate the factors of  $Z$ . The first few components of  $Z$  are consisted of significant information. The first few components are useful displaying the entire information. A correlation coefficient matrix is implemented to compute the coefficients and this matrix is numerically equal to the eigenvector of the correlation coefficient matrix [9].

v. Forecasting Engines: The computing model provides the choices of machine learning models and algorithms for either classification or regression. A set of algorithms such as SVM, KNN, ANN, and LSTM, are used for rainfall prediction. These models and algorithms are provided as forecasting engines in the prediction system.

vi. Result evaluation: Various evaluation metrics are used to evaluate classification and regression respectively. In classification, accuracy is most commonly used to measure the performance of classifiers, which can reveal a lot of predictive models' potential. In regression, the coefficients of determination ( $R^2$ ), mean square error (MSE), root mean square error (RMSE), and Pearson correlation coefficient (Pcc) are all adopted to measure the fitness of the regression model in the dataset. Once experimental results are received and saved to local storage, comparisons are made based on these evaluation metrics [10].

## II. LITERATURE REVIEW

Yajnaseni Dash, et.al (2017) suggested the variable monsoon trends over Kerala in winter, pre-monsoon, summer monsoon and post-monsoon periods [11]. Two AI (Artificial Intelligent) techniques namely SLFN (Single layer Feed-Forward Neural

Network) and ELM (Extreme Learning Machine) were adopted with the objective of predicting the rainfall of summer monsoon due to the occurrence of heavy rainfall in this period. The MAE (mean absolute error) acquired from the initial technique was calculated 6.39% and 3.87% from the latter technique. The results confirmed the superiority of ELM over other and this technique was utilized as a predictive tool to analyze the complex monsoon phenomena.

A. Kala, et.al (2018) introduced ANN (Artificial Neural Network) algorithm known as FFNN (Feed Forward Neural Network) to predict the rainfall [12]. This algorithm was effective and efficient soft computing technique using which rainfall was predicted. This algorithm was planned on the basis of on self-adaptive approach that allowed the model to learn from historical data containing functional relationships of data with predictive results on current data. To manage the water resources, it was essential to predict the rainfall accurately. The confusion matrix and RMSE (root mean square error) metrics were considered to evaluate the accuracy for predicting the rainfall. The results demonstrated that the introduced algorithm yielded higher accuracy.

Mary N. Ahuna, et.al (2017) projected a framework in order to predict the rainfall rate 30 seconds ahead of time with the help of ANN (artificial neural network) [13]. Thereafter, the resultant rainfall rate was taken in account to determine a suitable fade counter-measure. The historical rainfall rate patterns were taken over Durban. The future rain rate was predicted instantly using the resultant model. The projected framework was computed with regard to RMSE (root mean square error). The outcomes depicted resultant errors found within acceptable values at diverse rain events.

Hiyam Abobaker Yousif Ahmed, et.al (2020) designed a MLR (Multiple Linear Regression) algorithm for forecasting the RCP (rate of precipitation) such as rainfall rate for Khartoum state [14]. The website of the National Climatic Data Center was taken to gather the data. The designed algorithm was executed using Python code in which ANN (Artificial Neural Network) algorithm was implemented. The average MSE (mean square error) of training data was compared with the test data to evaluate the designed algorithm. The obtained results indicated that the designed algorithm led to enhance the MSE up to 85%.

ImrusSalehin, et.al (2020) developed a framework in order to predict the amount of rainfall with the deployment of AI (artificial intelligence) and LSTM (Long Short-Term Memory) methods [15]. The memory sequence data was quantified and the prior data was computed quickly to predict the rainfall effectively. These factors were assisted in determining the amount of rainfall. Six metrics such as temperature, dew point, humidity, wind pressure, wind speed, and wind direction were taken in account for forecasting the rainfall. The results confirmed that the developed framework offered the accuracy of 76%. This framework also emphasized on a large dataset in long time weather to attain superior result.

Oswalt Manoj S, et.al (2020) established a model to predict the rainfall on the basis of convLSTM (convolutional long short-term memory) which was capable of predicting the rainfall on the basis of spatial-temporal patterns [16]. The S-SGD (Salp-stochastic gradient descent) algorithm, in which SSA (Salp swarm algorithm) was combined with SGD (stochastic gradient descent) algorithm, implemented to tune the weights of established model so that good accuracy was obtained while predicting the rainfall. The MapReduce model was utilized for developing this model to handle the big data efficiently. The analysis of results attained on database confirmed that the established model offered a lower MSE (mean square error).

Gunawansyah, et.al (2017) intended ENN (Ensemble Neural Network) in which ANN (Artificial Neural Network) was put together with GA (Genetic Algorithm) for optimizing and investigating the best weights and biases [17]. One hidden layer was utilized in ANN architecture and ENN performed well on diverse dataset. The results validated that the intended approach offered the accuracy of 66.02% to predict the rainfall in dry season, 79.7% in wet season and 84.6% for all data scenario. Furthermore, it was useful to detect the anomaly rainfall so that the risk of loss was alleviated to start the planting time and growing season.

Sankhadeep Chatterjee, et.al (2018) recommended a new technique named HNN (Hybrid Neural Network) to predict the rainfall [18]. The feature set was diminished and the most promising attributes were discovered to predict the rainfall using Greedy Forward Selection algorithm. Initially, K-Means algorithm was implemented to cluster the data. Subsequently, the training of individual NN (Neural Network) was done for every cluster. A comparative analysis was conducted on the recommended technique against the traditional algorithm with regard to diverse parameters. The experimental results proved the supremacy of the recommended technique and this technique offered the accuracy of 84.26% when the technique of selecting feature was not utilized and 89.54% with feature selection to predict the rainfall.

### III. RESEARCH METHODOLOGY

The key motive of this work is to predict floods which has four phases. Various steps of the new approach are explained below:

1. **Pre-Processing:** - The dataset obtained the UCI depository is given as input in the initial phase of this research. Any kinds of missing values are eradicated and the data is cleaned in this the first phase. This is the crucial process to arrange data in the finite order. To remove missing values from the dataset, the mean of the dataset is calculated and missing values are removed with the mean value. The second stage of pre-processing is to define attribute and target set. In the dataset last column is the target set and rest are attribute sets. The dataset is treated in the data frames and when the dataset is treated in the data frame then, data will be processed with the column name. The third stage of the pre-processing is feature extraction for the further processing. In the next step of

feature extraction, the relationship is created amid each feat and target set. The relationship establishment process will find the most suitable attribute from the dataset. The most suitable attribute means the attribute on which most of the data depends. The most reliable data will be further processed for the classification.

**2. Selection of Multiple Classifiers:** -In this research work, an approach named voting classifier is used for predictive analysis. The three classifiers are selected which are given as input to the voting process. These classifiers include NB, KNN and Decision tree. K Nearest Neighbors (KNN) algorithm is a kind of supervised ML algorithm. The classification and regression predictive problems are done using this algorithm. This algorithm considers as lazy algorithm as there is not any specialized stage for training available in it. For the training, all the data has employed in classification. K-nearest neighbors is also known as non-parametric learning algorithm as it is unable to presume anything related to the fundamental data. The feature similarity has employed in KNN algorithm so as the new data points values are predicted. It is indicated that a value will be allocated by the new data point. This value is selected on the basis of the close matching of this value with the points in the training set. NB as a probabilistic classifier is commonly utilized to do classification. The working of this classification model is inspired from Bayes theorem which can be stated as:

$$P(A \setminus B) = \frac{P(B \setminus A) P(A)}{P(B)}$$

This theorem is used to find the probability of A happening, when B has happened. In the above theorem, B refers to the evidence while A denotes the hypothesis. This classifier makes assumption about the predictors/features being independent. This means that the occurrence of one specific feature does not affect the other feature. This is the reason that this algorithm is referred as naive. Decision tree is a very

popular approach which is generally used for classification and prediction. The configuration of Decision tree is just like a tree. In this configuration, each internal node refers to a test on an attribute. All branches of this tree represent the tested outcomes. Also, a class label is given to every leaf node. This node is referred as terminal node. The partitioning of source set is performed into subsets on the basis of a feature value test for tree learning. This process is repeated again and again on all resultant subsets. This process is referred as recursive partitioning. The recursion is finished after the subset at a node get the value same as the target variable, or when partitioning no more inserts value to the forecasting. There is no need of any field info or metric set-up for generating this classifier. Therefore, this classifier simplifies the analysis process of knowledge discovery. This classification model can manage huge volume of data. This classifier generally produces very accurate outcomes.

**3. Voting Method:** -The voting classification is the final process of the prediction analysis. In the process of voting, three classifications which are decided in the step 3 are taken as input for the voting. These classifiers include NB, KNN and decision tree. The voting classification is broadly classified into hard and soft voting method. The soft voting classification method is used for the predictive analysis. In the method of voting, all the three classifiers are taken as input and weight of each classifier will be defined. The weight defines the priority of each classifier according to the classifier accuracy. The whole dataset is separated into two subsets of training and testing respectively. The first subset comprises 60% while the second subset covers 40% of the whole data. The voting classifier will train the model and test the model based on the training set. The efficiency of devised approach will be examined w.r.t accuracy, precision and recall.

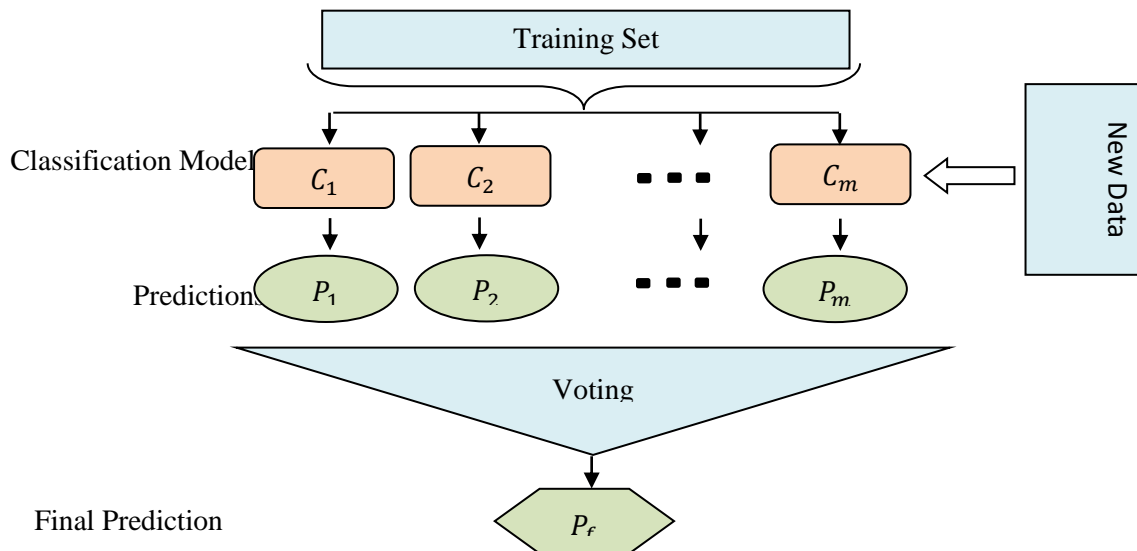


Fig 2: Proposed Methodology

IV. RESULT AND DISCUSSION

The dataset for the work is taken from the UCI database. This work considers two cases. In the first case, the use of SVM

approach is carried out while the second new case applies voting classifier for predicting rainfall. In this case, the efficiency test of the new scheme is also performed.

**Table 1:** Dataset Description

<b>Data Set Characteristics:</b>	Multivariate, Time-Series	<b>Number of Instances:</b>	N/A	<b>Area:</b>	Life
<b>Attribute Characteristics:</b>	Categorical, Integer	<b>Number of Attributes:</b>	20	<b>Date Donated</b>	N/A
<b>Associated Tasks:</b>	N/A	<b>Missing Values?</b>	N/A	<b>Number of Web Hits:</b>	367821

Following is the list of performance metrics -

1. Precision: Precision refers to the ratio of the number of related records retrieved to the total number of unrelated and related records retrieved.

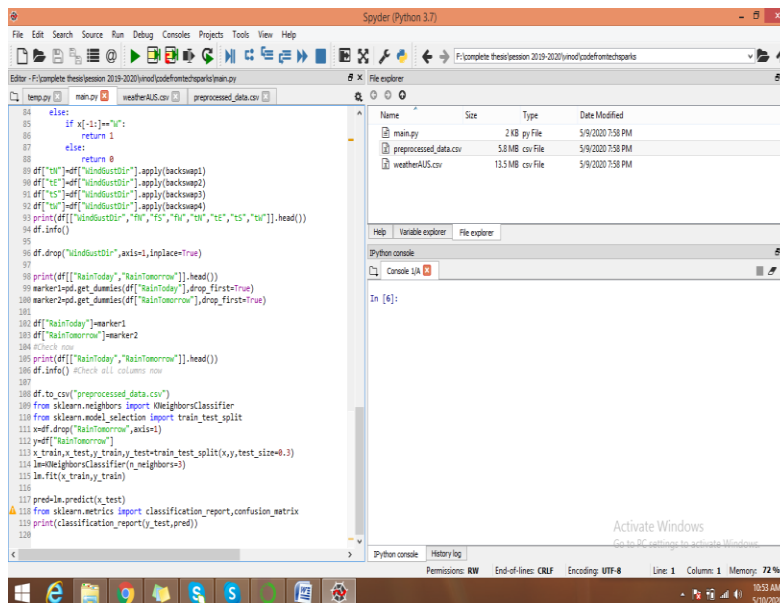
$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

2. Recall: This parameter represents the ratio of the number of related records retrieved to the total number of related records in the database.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

3. Accuracy: It is the ratio of total number of points classified suitably to the total number of points multiplied by 100.

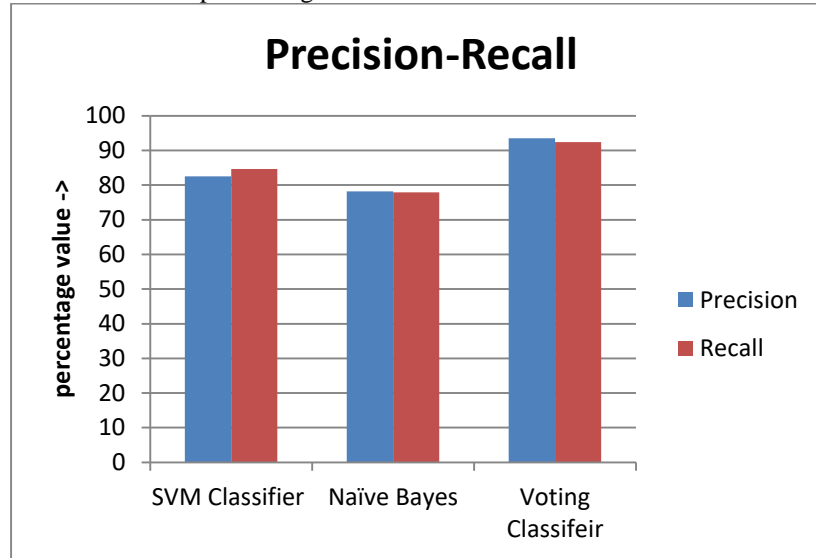
$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$



**Fig 3:** Apply Voting Classifier

Figure 3 shows the use of dataset as input for the rainfall prediction. The dataset has various attributes. The processing of these attributes is carried out for predictive analysis. This dataset will be pre-processed for the further processing. The

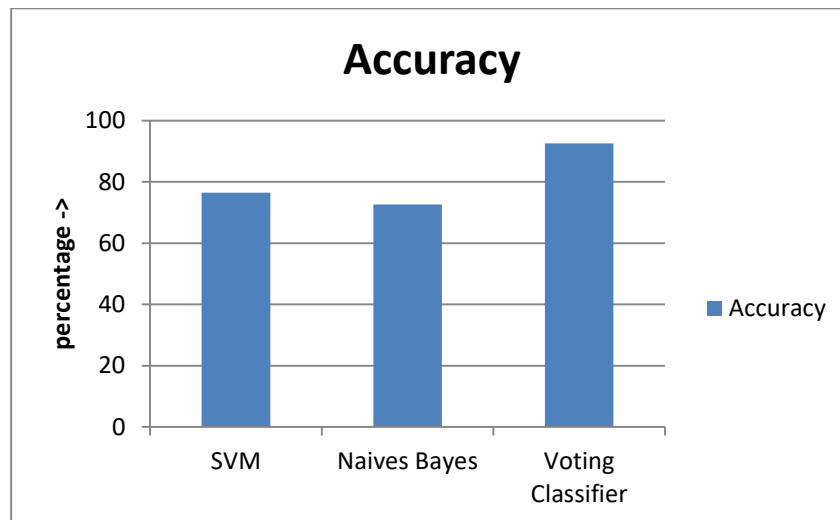
pre-processing removes missing and unnecessary values from the dataset. This figure also shows the implementation of voting classifier for predicting rainfall.



**Fig 4:** Precision-recall comparison

Figure 4 shows the comparison between three classifiers (SVM, NB and voting) w.r.t precision-recall values. The

voting classifier as an ensemble classifier achieves maximal value of precision-recall among other two classifiers.



**Fig 5:** Accuracy comparison

Figure 5 shows the comparison between three classifiers (SVM, NB and voting) w.r.t accuracy. The voting classifier as an ensemble classifier achieves highest accuracy rate among other two classifiers.

## V. CONCLUSION

In the last few years, researchers have developed and practiced many weather forecasting models using data mining methodologies. These forecasting models have shown great

accuracy in weather prediction. Precipitation prediction with data mining technology, which is different from classic techniques of weather forecasting, has drawn a lot of attention from the research community. Based on machine learning theory, historical observational data can be used to predict future rainfall. Contrary to other types of models, this computing operation is clearly more suitable. There are two stages included in this task i.e. training and testing. In the first stage, the training of rainfall prediction system is carried out

for a classification model. The voting classifier will be applied for the rainfall prediction. The voting classifier integrates three classifiers which are KNN, NB and DT for the prediction analysis. The implementation of new algorithm is carried out in python software and outcomes are analyzed w.r.t accuracy, precision, and recall. The analytic outcomes reveal that accuracy, precision and recall of the proposed model is highest among the other two models.

## VI. REFERENCES

- [1] J. Niu and W. Zhang, "Comparative analysis of statistical models in rainfall prediction," 2015 IEEE International Conference on Information and Automation, 2015, pp. 2187-2190
- [2] N. Sethi, Dr. K. Garg, "Exploiting Data Mining Technique for Rainfall Prediction", 2014 (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, PP. 3982-3984
- [3] A. Geetha and G. M. Nasira, "Data mining for meteorological applications: Decision trees for modeling rainfall prediction," 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014, pp. 1-4
- [4] I. Wahyuni, W. F. Mahmudy and A. Iriany, "Rainfall prediction in Tengger region Indonesia using Tsukamoto fuzzy inference system," 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2016, pp. 130-135
- [5] A. H. Manek and P. K. Singh, "Comparative study of neural network architectures for rainfall prediction," 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), 2016, pp. 171-174
- [6] N. Hasan, N. C. Nath and R. I. Rasel, "A support vector regression model for forecasting rainfall," 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT), 2015, pp. 554-559
- [7] A. Kusiak, X. Wei, A. P. Verma and E. Roz, "Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach," in IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 4, pp. 2337-2342, April 2013
- [8] C. P. Shabariram, K. E. Kannammal and T. Manojpraphakar, "Rainfall analysis and rainstorm prediction using MapReduce Framework," 2016 International Conference on Computer Communication and Informatics (ICCCI), 2016, pp. 1-4
- [9] F. Nhita and Adiwijaya, "A rainfall forecasting using fuzzy system based on genetic algorithm," 2013 International

Conference of Information and Communication Technology (ICoICT), 2013, pp. 111-115

- [10] V. B. Nikam and B. B. Meshram, "Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach," 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation, 2013, pp. 132-136
- [11] Yajnaseni Dash, S.K. Mishra, B.K. Panigrahi, "Rainfall prediction of a maritime state (Kerala), India using SLFN and ELM techniques", 2017, International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICT)
- [12] A. Kala, S.Ganesh Vaidyanathan, "Prediction of Rainfall Using Artificial Neural Network", 2018, International Conference on Inventive Research in Computing Applications (ICIRCA)
- [13] Mary N. Ahuna, Thomas J. Afullo, Akintunde A. Alonge, "Rainfall rate prediction based on artificial neural networks for rain fade mitigation over earth-satellite link", 2017, IEEE AFRICON
- [14] HiyamAbobaker Yousif Ahmed, Sondos W. A. Mohamed, "Rainfall Prediction using Multiple Linear Regressions Model", 2020, International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCEEE)
- [15] ImrusSalehin, Iftakhar Mohammad Talha, Md. Mehedi Hasan, Sadia Tamim Dip, Mohd. Saifuzzaman, Nazmun Nessa Moon, "An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network", 2020, IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)
- [16] Oswalt Manoj S, Ananth J P, "MapReduce and Optimized Deep Network for Rainfall Prediction in Agriculture", 2020, The Computer Journal
- [17] Gunawansyah, Thee HouwLiong, Adiwijaya, "Prediction and anomaly detection of rainfall using evolving neural network to support planting calender in soreang (Bandung)", 2017, 5th International Conference on Information and Communication Technology (ICoICT)
- [18] Sankhadeep Chatterjee, Bimal Datta, Soumya Sen, Nilanjan Dey, Narayan C. Debnath, "Rainfall prediction using hybrid neural network approach", 2018, 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)