Performance Comparison of Machine Learning Algorithms in Identifying Dry and Wet Spells of Indian Monsoon

Harikumar Rajaguru and S. R. Sannasi Chakravarthy

Abstract For water-related industries, the characteristics of wet spells and intervening dry spells are highly useful. In the face of global climate change and climate-change scenario forecasts, the facts become even more important. The goal of this study is to determine the wet and dry spells that occur throughout the monsoon season in peninsular India. The India Meteorological Department (IMD) observations were made over the course of a hundred days, from October 23 to January 30, 2019, with 334 rainy days and 60 dry days. The IMD data provides ten observational characteristics in peninsular India, including maximum, minimum, and average temperatures, rainfall wind speed, atmospheric pressure, illumination, visibility, relative cloud density, and relative humidity. Four statistical factors, such as mean, variance, skewness, and kurtosis, further decrease these characteristics. The observed characteristics and their statistical parameters follow a nonlinear trend, as seen by histogram plots. For assessing the classification performance, a collection of four algorithms is used: Logistic regression, gradient boosting, Gaussian mixture model, and firefly with Gaussian mixture model. During both the dry and rainy spells of monsoon observation, all of the classifiers achieve greater than 85% classification accuracy (average).

Keywords Dry spell · Wet spell · Monsoon · Water scarcity · Machine learning · Firefly

1 Introduction

Rainfall occurs in spells in tropical monsoonal regions and is a seasonal phenomena [1]. Classifiers and other criteria are used to describe the start and conclusion of the rainy season, as well as the frequency, quantity, and intensity of rainfall, the duration of wet spells (WSs), and the duration of intervening (between two rain spells) dry spells (DSs). Weather forecasting is a work that uses science and technology to

H. Rajaguru (✉) · S. R. Sannasi Chakravarthy
Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam 638401, India

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 336, https://doi.org/10.1007/978-981-16-6723-7_19

anticipate the state of the atmosphere for a certain time and location in the future [1]. Since ancient times, humans have sought to forecast the weather. Rainfall prediction is one of the most essential aspects of weather forecasting, since it is crucial for food production, water resource management, and many other outdoor activities. The problem's most important difficulty is determining the rainy season's annual beginning and ending dates, as well as identifying wet and dry periods in the rainfall time distribution [2]. Using classifiers and parameters, we attempt to determine the wet and dry dates in a given monsoon season in this work. The India Meteorological Department (IMD) observations in this study were obtained during a period of one hundred days, from October 23 to January 30, 2019, with 334 rainy days and 60 dry days. From the IMD data for peninsular India, ten observational characteristics such as maximum, minimum, and average temperatures, rain fall wind speed, atmospheric pressure, illumination, visibility, relative cloud density, and relative humidity are obtained with the label of wet and dry spell. Using classifiers and input parameters, the number of wet and dry spells is calculated. The outcomes are then compared with the findings of IMD labels.

The workflow proposed for the research is depicted in Fig. 1. From this figure, the database is visually analyzed using graphs and pre-processed for their better results and the data classification is then implemented through the four distinct ML algorithms, namely logistic regression, Gradient boosting, Gaussian mixture model, and firefly with Gaussian mixture model classifiers. As a final point, the comparison of results is done for the performance of classifying dry and wet spells.

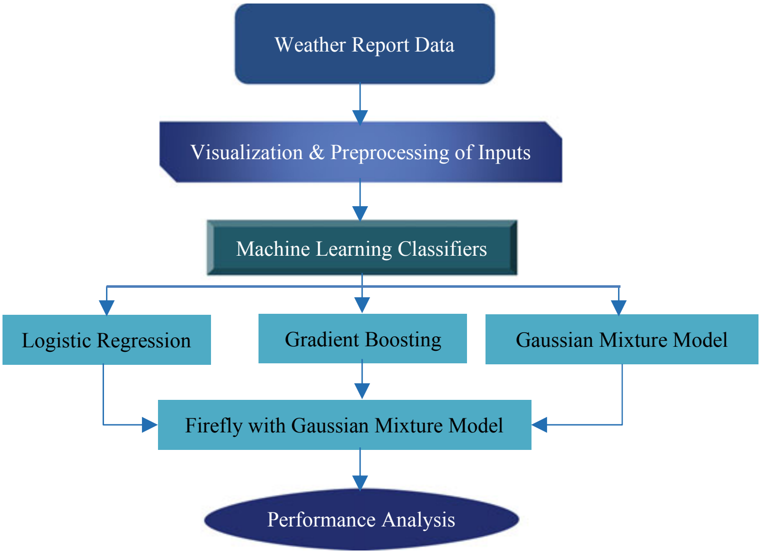
Weather Report Data
Visualization & Preprocessing of Inputs
Machine Learning Classifiers
Logistic Regression
Gradient Boosting
Gaussian Mixture Model
Firefly with Gaussian Mixture Model
Performance Analysis

Fig. 1 Workflow proposed

2 Materials and Methods

2.1 *Input Dataset*

From the India Meteorological Department (IMD) data that acquired for peninsular India, includes the output classes of dry and wet spells during the season of monsoon. This has the ten observational attributes such as max, min, and average temperature values, speed of rainfall wind, pressure of atmosphere, relative cloud-density and humidity, visibility values, and illumination values. In the input dataset, the total number of rainy days is 334, while the total number of dry days is 60. The number of rainy days outnumbers the number of sunny days. As a result, it is collected during the rainy season.

2.2 *Data Visualization of Input Data*

Irrespective of classification problem solving through ML algorithms, the analysis of data input is very crucial for further research phases. The univariate input data analysis through a distribution plot has carried out and is given in Fig. 2.

From Fig. 2, the average temperature and humidity attribute values are inferred as a much right-skewed one in the input dataset. In addition, it reveals that the

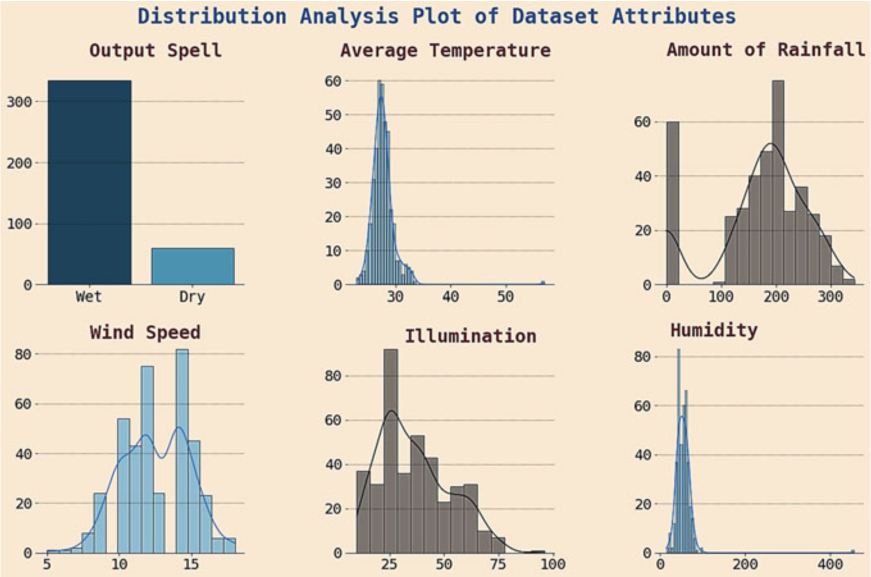

Fig. 2 Univariate input analysis through distribution plot

dataset includes more rainfall intensity values for rainy spell in the range of 100–350%, and the wind speed is equally distributed throughout the season. Moreover, the illumination values of input data seem to be substantially skewed. The above discussion implies that the input data needs to be normalized before classifying the data.

2.3 Data Preprocessing

As from the input dataset, the distribution plot of Fig. 2 provides that the input measurements are being highly nonlinear and overlapped. Also, while inspecting each attributes of the input data, it is found that the data needs to be normalized before data classification. And so, the input measurements are normalized then though the use of Standard Scalar technique [3]. As a result, the data input is now ready for next succeeding step of data classification as depicted in Fig. 1.

3 Classification Algorithms

The paper employs a hybridized algorithm that includes the advantage of Gaussian mixture model concept with the nature inspired firefly algorithm. Also, the base classifiers, namely logistic regression, Gradient boosting, and Gaussian mixture model algorithms are employed. The algorithms of the above-said classifiers are detailed in this section.

3.1 Classification Using Logistic Regression (LR) Algorithm

In this type of logistic regression means of classification, statistical methods are adopted for predicting the binary targets that includes rainy wet and non-rainy dry spells [4]. The LR algorithm being a linear learning technique, it generally make use of the odds of an event for performing predictions with logistic regression concept. For this action, the LR approach employs a simple sigmoidal mathematical function for mapping of all input data points to their binary targets [5]. As a result, an S-shaped curve can be represented as the note of traditional logistic function. This could be depicted mathematically using a simple sigmoidal equation as shown in below equation of [5],

$$\text{Sigmoidal Function} = \frac{1}{1 + \mathrm{e}^{-x}} \tag{1}$$

3.2 Gradient Boosting (GB) Classifier

The Gradient boosting algorithm is a simple collection of ML models that includes several weak-learning algorithms for building a powerful prediction classifier [6]. While implementing Gradient boosting, decision trees are commonly utilized. The Gradient boosting models are gaining popularity as a result of their ability to categorize complicated information of input dataset [6]. The decision tree (DT)-based GB is employed in this paper, where the implementation steps are summarized below [7],

Step 1: Computing the average value of the output binary targets.
Step 2: Computing the residual values computed as a difference of actual and prediction.
Step 3: Construction of DT is done.
Step 4: Predicting the output binary target by the use of every trees created in the ensemble.
Step 5: Repeat the computation of new residual values.
Step 6: Repeating the steps of 3 to 5 with the condition of matching the number of iterations with the amount of estimators used.
Step 7: Once completion of training, make use of all the trees in the ensemble for making a conclusion on final prediction as one of the output targets.

3.3 Gaussian Mixture Model (GMM) Classifier

Gaussian mixture model (GMM), probability-based algorithm used more common for depicting normally distributed sub population over overall populations [8]. The GMM algorithm is actually utilized for unsupervised learning problems for learning the sub-population and the automatic assignment of sub populations. However, in this paper, the GMM algorithm is employed for classification or supervised learning problems for learning the boundaries of sub population. After the training phase, that is, once fitting the data with GMM, it can classify which of the cluster a newer data point belongs to. But, this is possible only if the GMM is provided with the target labels. Here, it is very important that the clusters are chosen arbitrarily, and its probability density function can be defined as [9],

$$y = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

where μ and σ represent the mean and standard deviation of the input data. Here, the probability of an input data point can be calculated as [9],

$$p(x) = \sum_{i=1}^{k} \varphi_i \aleph(x|\mu_i, \sigma_i) \tag{3}$$

3.4 Firefly with GMM as a Classifier

Unlike using GMM as an unsupervised learning approach, the paper utilizes the GMM for solving supervised learning problem. However, the performance of the GMM would not be a satisfied one while comparing with other conventional ML algorithms. Thus, in order to improve their prediction ability in supervised learning problems, the paper hybridized the metaheuristic firefly algorithm with the Gaussian mixture model algorithm. This type of hybrid implementation involves in helping the prediction by removing the insignificant data points and outliers and so making the GMM model to provide better accuracy in supervised approaches. The parameters of firefly algorithm are selected as experimented in our previous work [10].

4 Results and Discussion

The research work implemented as depicted in Fig. 1 of this paper is done using Google Colab which is an online IDE research base provided by Google though a personal Gmail account used on the web browser, Google Chrome. The data inputs after preprocessing as illustrated in Sect. 2 have accordingly splitted for the phase of classification with 70:30 standard with 70% of training inputs and 30% of testing inputs. As depicted in Fig. 2 (first column plot), the input data comprises more wet sample class targets than dry class target, so there might be a problem of class imbalance. For overcoming this class imbalance problem, SMOTE type [11] of splitting data is used. The number of total input data taken and its split up for training and testing phase are portrayed in Fig. 3. In addition to this, prior to classification part of implementation, the input data processed can be normalized through min–max standardization [3] technique as per the equation shown below,

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

where $x_{\min}, x_{\max}$ denote the minimum and maximum of data point values and x refers to the input vectors.

As discussed above, after normalization and splitting of preprocessed data, the considered ML algorithms are then fitted (trained) and tested to check the efficacy in predicting dry and wet spell. That is, the LR, GB, and GMM classifiers together with their hybridized classifier model, i.e., firefly with GMM algorithms are employed for

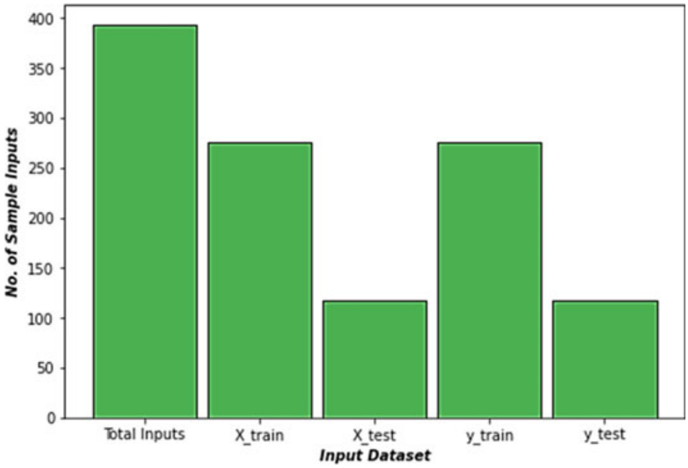

Fig. 3 Splitting of data inputs for classification

this prediction. In addition, the performance is fivefold cross-validated to provide better results. The results obtained in classifying spells can be assessed using the benchmark measures [12] which are generally a standard one in the problems of binary prediction. The metrics adopted in the paper are sensitivity, accuracy, specificity, F1 score, and precision. Here, the above-said performance measures are derived or taken through the confusion matrix (CM) which comprises the particulars of true and false negatives and positives. These results are then validated through a standard measure, Matthews Correlation coefficient (MCC).

The obtained elements of CM regarding each classification algorithms are graphically plotted in Fig. 4. It is noted from this graph that the amount of true negative

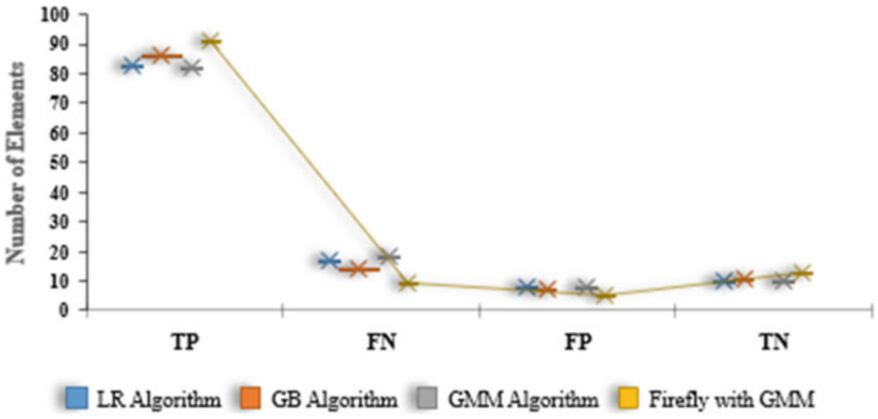

Fig. 4 Plot of confusion matrix of ML algorithms

Table 1 Performance of algorithms used for classification

Classifiers	Performance comparison (%)					
	Sensitivity	Specificity	Accuracy	Precision	F1 score	MCC
LR classifier	83	55.56	78.81	91.21	86.91	33
GB classifier	86	61.11	82.2	92.47	89.12	41.45
GMM model	82	55.56	77.97	91.11	86.32	31.74
Firefly with GMM classifier	91	72.22	88.14	94.79	92.86	58.37

and positive elements (TN and TP) is high for the Gradient Boosting classification algorithm as compared with the LR classifier. And while including the performance of the GMM classifer, still the true negative and positive elements of CM is high for the Gradient Boosting algorithm. As implemented with the hybrid algorithm for classification that includes the firefly algorithm together with the GMM model, the true prediction elements of CM significantly improved as illustrated in the plot of Fig. 4. In a similar way, while considering the false misclassification, the GMM model as a classifier provides more FN and FP elements as compared with other base classifiers. However, the same GMM classifier together with the Firefly algorithm provides very less misclassification in this prediction problem. Moreover, the prediction, i.e., the amount of false classification gets depreciated and so the amount of correct predictions gets improved for the Firefly with GMM model as depicted in Fig. 4. The discussion on the results obtained using performance metrics will be further discussed in detail.

The performance obtained using the above-said classification algorithms are listed in Table 1. In this table, the logistic regression algorithm's performance as compared with the GMM algorithm is higher. But the Gradient boosting algorithm provides a better performance than this logistic regression classifier. This implies that the Gradient boosting classifier provides a better accuracy of 82.2% accuracy, 92.47% of precision, and 89.12% of F1 score. Here, it is noted that while comparing the individual base algorithms, the gradient boosting algorithm yields the high classification, and so, the value of MCC for GB classifier is attained as 41.45 which is supreme over other base classifiers. For further improving its performance, the hybridization technique is used by making use of the metaheuristic approach.

The classification performance of this hybrid firefly with GMM algorithm while comparing with other algorithms is plotted graphically in Fig. 5. In this graph, the hybridized firefly with GMM algorithm yield 88.14% of accuracy with a precision of 94.79%, and F1 score of 92.86%. These obtained performances are validated by the MCC attainment value of 58.37, and it is obviously higher than other employed classification algorithms. Hence, the hybridized firefly together with the GMM algorithm attains a maximum performance than the LR, GB, and GMM algorithms as depicted in Table 1 and Fig. 5.

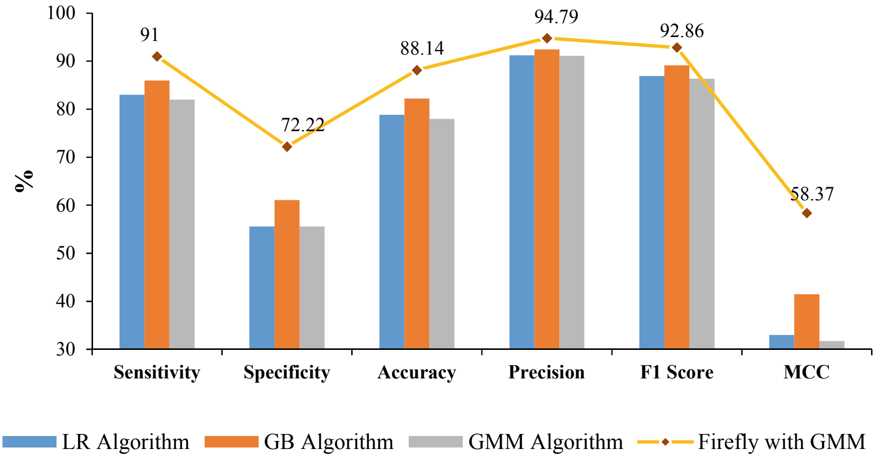

Fig. 5 Graphical comparison plot used for classifiers' performance

5 Conclusion and Future Scope

The work proposed a hybridized approach for developing a weather forecasting predictor used in the prediction of rainfall through several measurements. The algorithms used for classification are logistic regression, gradient boosting, and Gaussian mixture model. Here, the prediction performance has been improved further by incorporating the firefly algorithm together with the above-said Gaussian mixture model algorithm. For this evaluation, the dataset considered has ten different measurements with the inputs taken from 334 rainy days and 60 dry days. That is, the dataset comprises of 694 input samples taken from 694 different climatic days. And this input data is analyzed graphically using the distribution plot which revealed the nonlinearity nature of the inputs. Then, the input data is normalized and fed for different classifiers for prediction. For this, the input data had been splitted using a standard ratio of 70:30 by means SMOTE technique. As the aim of the research, the work attains a supreme performance of 88.14% accuracy with the improved value of MCC as 58.37. Several algorithms and approaches are proposed for efficient rainfall prediction are now available in literature, but there is still a need for a comprehensive literature review and systematic mapping research that can represent proposed solutions, current challenges, and current developments in this sector. The outcome of this research add to the existing body of knowledge in numerous ways. For engineers, scientists, managers, and planners working in water-related industries, the climatology and variability of the rainy season's characteristics, as well as wet and dry periods, are invaluable information. The focus of future study will be on using other metaheuristic algorithms with different preprocessing techniques.

References

1. B.T. Pham, L.M. Le, T.T. Le, K.T.T. Bui, V.M. Le, H.B. Ly, I. Prakash, Development of advanced artificial intelligence models for daily rainfall prediction. Atmos. Res. **237**, 104845 (2020)
2. N. Mishra, H.K. Soni, S. Sharma, A.K. Upadhyay, Development and analysis of artificial neural network models for rainfall prediction by using time-series data. Int. J. Intell. Syst. Appl. **10**(1) (2018)
3. C. Abirami, R. Harikumar, S.S. Chakravarthy, in *Performance Analysis and Detection of Micro Calcification in Digital Mammograms Using Wavelet Features*. 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (IEEE, 2016), pp. 2327–2331
4. A. De Caigny, K. Coussement, K.W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. Eur. J. Oper. Res. **269**(2), 760–772 (2018)
5. T. Pranckevičius, V. Marcinkevičius, Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic J. Mod. Comput. **5**(2), 221 (2017)
6. S. Georganos, T. Grippa, S. Vanhuysse, M. Lennert, M. Shimoni, E. Wolff, Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. IEEE Geosci. Remote Sens. Lett. **15**(4), 607–611 (2018)
7. F. Climent, A. Momparler, P. Carmona, Anticipating bank distress in the Eurozone: an extreme gradient boosting approach. J. Bus. Res. **101**, 885–896 (2019)
8. A. Das, U.R. Acharya, S.S. Panda, S. Sabut, Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques. Cogn. Syst. Res. **54**, 165–175 (2019)
9. Y. Li, W. Cui, M. Luo, K. Li, L. Wang, Epileptic seizure detection based on time-frequency images of EEG signals using Gaussian mixture model and gray level co-occurrence matrix features. Int. J. Neural Syst. **28**(07), 1850003 (2018)
10. S.R. Sannasi Chakravarthy, H. Rajaguru, Detection and classification of microcalcification from digital mammograms with firefly algorithm, extreme learning machine and non-linear regression models: a comparison. Int. J. Imaging Syst. Technol. **30**(1), 126–146 (2020)
11. S.R. Sannasi Chakravarthy, H. Rajaguru, A novel improved crow-search algorithm to classify the severity in digital mammograms. Int. J. Imaging Syst. Technol. **31**, 921–954 (2021). https://doi.org/10.1002/ima.22493
12. S.R. Sannasi Chakravarthy, H. Rajaguru, Lung cancer detection using probabilistic neural network with modified crow-search algorithm. Asian Pac. J. Cancer Prev. APJCP **20**(7), 2159 (2019)