Deep Learning Based LSTM and SeqToSeq Models to Detect Monsoon Spells of India[*]

Saicharan Viswanath[1], Moumita Saha[2], Pabitra Mitra[3], and Ravi S. Nanjundiah[2,4,5]

[1] National Institute of Technology Karnataka, Surathkal, India
[2] Centre for Atmospheric and Oceanic Sciences, Indian Institute of Science, Bangalore, India
[3] Indian Institute of Technology Kharagpur, India
[4] Divecha Centre for Climate Change, Indian Institute of Science, Bangalore, India
[5] Indian Institute of Tropical Meteorology, Pune, India
{vsaicharan1998,moumita.saha2012,pabitra,ravisn}@gmail.com

Abstract. Monsoon spells are important climatic phenomenon modulating the quality and quantity of monsoon over a year. India being an agricultural country, identification of monsoon spells is extremely important to plan agricultural policies following the phases of monsoon to attain maximum productivity. Monsoon spells' detection involve analyzing and predicting monsoon at daily levels which make it more challenging as daily-variability is higher as compared to monsoon over a month or an year. In this article, deep-learning based long short-term memory and sequence-to-sequence models are utilized to classify monsoon days, which are finally assembled to detect the spells. Dry and wet days are classified with precision of 0.95 and 0.87, respectively. Break spells are observed to be forecast with higher accuracy than the active spells. Additionally, sequence-to-sequence model is noted to perform superior to that of long-short term memory model. The proposed models also outperform traditional classification models for monsoon spell detection.

Keywords: Active spell · Break spell · Long short term memory · Sequence-to-sequence model · Attention mechanism · Classification · Spells' detection

1 Introduction

Indian monsoon is mainly governed by south-west monsoon, occurring between June and September. The quality of monsoon during this period is extremely important for growth and prosperity of the country. The summer rainfall not only regulates the agricultural productivity but also assists in fresh water renewal, and nourishes the flora and fauna of the subcontinent. The characteristics and variability of Indian monsoon are widely studied in literature owing to its importance to the country [10, 6, 12, 11, 4]. Monsoon variability is high and its prediction resembles a provocative area in the field of climate research. The distribution of rainfall over the country on spatial basis as well as the temporal basis during the monsoon period are non-uniform. There can be continues period of days when it is completely dry or it can be heavily raining for a stretch

---

of days. The monsoon process bears high variability and uncertainty. Dry days can lead to low monsoon year or even a drought year. Similarly, continues wet days with high rainfall values can lead to a excess rainfall year or even a disastrous flood. Thus, classifying monsoon on daily scale at a lead is extremely important to frame the policies and prevention measures for the country.

The article aims to identify the spells during monsoon months. Monsoon spells are defined by meteorologist considering various climatic properties like cloud nature and properties, different synoptic conditions, rainfall anomalies, position of monsoon trough, and wind strength and directions [19, 9, 5]. Monsoon spells are categorized into two phases– break and active spells. Break spells are defined as minimum three days or more at continuum having no rainfall or standard deviation below the mean rainfall. On a similar note, active spells refer to three or more days at continuum when the rainfall amount is at least standard deviation above the mean rainfall. Rajeevan et al. [18] proposed the standards for monsoon spells, characterized them to establish that the active spells are of less duration than the break spells. Active, weak, and break spells are also studied utilizing the horizontal wind shear [16]. Kumar and Dessai [13] have worked towards identifying breaks of Indian summer monsoon and showed their results consistent with conventional methods. Mandke et al. [14] have highlighted the variation of break and active spells with enhanced carbon-dioxide concentration, using global climate models. Ensemble models of National Centre for Environmental Prediction (NCEP) are also explored to forecast the monsoon spells of India [1]. Linear discriminant analysis seems to assists in the prediction of spells for Indian sub-continent [25]. Rao et al. [20] have studied the climatological characteristics of rainfall over India during break and active spells. The future projection of Indian monsoon's break and active spells are simulated by Sudeepkumar et al. [28] and characterized for different regions considering the climate change scenario.

Deep-learning based approach is proposed to detect the monsoon spells of India. Data-driven machine-learning or deep-learning methods have shown their efficiency in different climate problems owing to availability of huge climatic data [3, 7, 15, 23, 21, 22]. Deep-learning based autoencoder and stacked autoencoder have been used to predict aggregate Indian monsoon [24], regional Indian monsoon [27] at annual scale with high accuracy. Deep learning method is also utilized to predict early and late temporal phases of monsoon [26]. In this article, we propose identification of break and active monsoon spells of India using long short-term memory (LSTM) and sequence-to-sequence (Seq2Seq) models. Both LSTM and Seq2Seq models are different flavors of recurrent neural network (RNN). The prime reason for choosing recurrent neural networks is that they are capable of capturing the temporal variation in data, which is essential for time series prediction. LSTM networks uses special units in addition to standard RNN units, which help them to remember or forget information over long sequences (as required for modeling). Identification of spell is modeled as a classification problem, where we first attempt to classify each day as dry, wet, or normal day. Finally, the classification results at daily scale are summed up to discover the break or active monsoon spells. Spells are detected at a lead of one day to five days.

The continuing article is ordered as follows. The input data-set, output rainfall classes, and the detailed preprocessing steps are discussed in Section 2. Section 3 elaborates

the proposed approach with brief description of LSTM and Seq2Seq models, and various experimental set-ups. Section 4 highlights the experimental outcomes and analysis, along with the performance metrics used. Lastly, Section 5 presents the conclusion and future scopes of the work.

2   Data and Preprocessing

2.1   Data Sources

Monsoon spell detection involves considering daily rainfall during June-September for the central region of Indian subcontinent, which extends the geographical lattitide of $21\,°$N - $27\,°$N, and longitude of $72\,°$E - $85\,°$E. India Meteorological Department provides the daily rainfall at $1\,°$ x $1\,°$ spatial resolutions [17]. Four climatic variables, namely, sea level pressure (SLP), air temperature (AT), v-wind (VWND), and u-wind (UWND) are considered as input variable to classify the monsoon days. All the input variables are considered at surface and the daily values are collected from reanalysis data prepared by NCAR [8], provided at $2.5\,°$ x $2.5\,°$ spatial resolutions for central India. All the data are collected for 1948–2014.

2.2   Preprocessing: Normalizing Features

All the input variables and rainfall data are normalized. Normalization is performed with removing the mean, followed by dividing the residual with the standard deviation of data (Equation 1). Both the mean and standard deviation are computed considering the training data. This is done to prevent leaking of information from validation and test sets. Normalizing data helps in avoiding training problems like local optima and weight decay. It also leads to faster convergence to global optima.

$$X_{\text{norm}} \leftarrow \frac{X - Mean(X_{\text{train}})}{Std(X_{\text{train}})} \tag{1}$$

2.3   Prepropessing: Daily Rainfall Class Assignment

After normalizing the rainfall to mean zero and unit standard-deviation, daily rainfall are labeled among three rainfall classes– dry, wet or normal. A sample or day is labeled as dry class if the normalized rainfall is below -1, and it is labeled as wet class if the normalized rainfall is above +1. Finally, days or samples which receive rainfall having rainfall in middle of these two extremes are marked as normal class. These class labels are used further for classification task.

3   Methodology

The proposed method for predicting the break and active monsoon spells of India consists of the following steps– (a) flattening the spatial-temporal input data, (b) classifying the days into dry, wet, or normal class using LSTM or Seq2Seq models, (c) Summing up the classified days to detect break or active monsoon spells. The summary of the proposed approach to detect the monsoon spells (break or active) is shown in Figure 1.
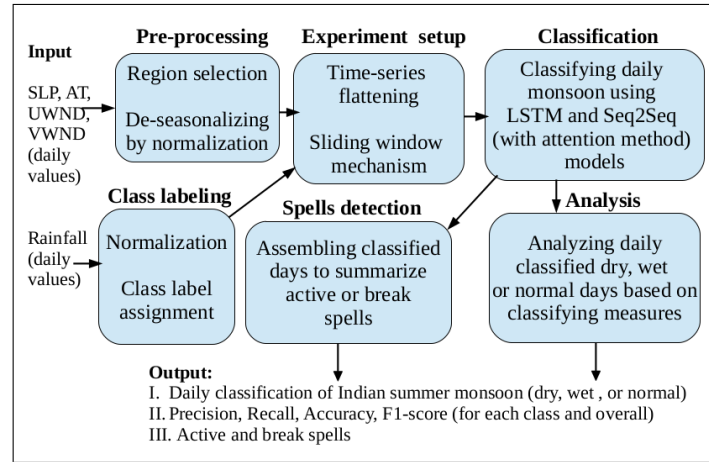
Fig. 1: Block-flow of the proposed method for identifying break and active monsoon spells of India with LSTM and Seq2Seq models

3.1   Input Time Series Flattening

The input consists of dimension equivalent to $TimeStepHist \times NumFeatures$, where each row or tuple represents the value of input variables on a particular day and each column represents the distinct input variable. As an example, if we are using rainfall, AT and SLP of four previous days history (day 1 to day 4) as our input to classify the rainfall of succeeding day (day 5), then our input will have dimension 4×3 , where four indicates the number of past history of days used ($TimeStepHist$) and three indicates the number of input features ($NumFeatures$). Considering all the examples of rainfall classification tuples lead to adding a third dimension to the input ($NumExamples \times TimeStepHist \times NumFeatures$). However, for feeding the input to the proposed model, it is required to flatten the input tensor into two-dimension matrix, which is arranged in form of $NumExamples \times (TimeStepHist \times NumFeatures)$ (i.e. the last two dimension are flattened to single dimension as feature1 at timehist1, feature2 at timehist1,......,feature1 at timehist2, feature2 at timehist2, and so on).

3.2   Classification using Long Short Term Memory Network Model and Sequence-to-Sequence Model

Classification of the rainfall days during June-September at a lead is performed using two mutants of recurrent neural network, namely, sequence-to-sequence and long short term memory models.

Long short term memory model: Long short term memory networks (LSTM) is an extension and flavor of recurrent neural network (RNN), used for learning long-term dependencies. RNN consists of feedback connection between the layers. The motivation behind using this network lies in presence of sequential observations, where it is

required to forecast future time-series by learning from the past time-steps. The value of the climate variables for a certain time-step may get influenced by the same variable values at the past. This signify the use of RNN architectures for predicting spells of Indian monsoon. The hidden layers act as a memory unit by storing the information captured from previous states of the sequential input. The hidden layer of RNN can be described as in Equation 2.

$$\text{hid}_i = f(\text{RecWeight}_{\text{hid}}\text{h}_{i-1} + \text{Weight}_{\text{hid}}x_i + \text{bias}_{\text{hid}}) \tag{2}$$

The function f corresponds to the activation function for introducing non-linearities (eg. sigmoid, tanh, etc.), and the mapping from hidden to the output layer is mainly considered as softmax for classification. Despite of all its capabilities, RNN are prone to the *vanishing gradient problem*, and hence unable to learn long term dependencies. On the other hand, LSTM can learn and adapt the existing long distance dependencies. They are capable of remembering the useful information over a longer sequences and forgets the other through special structures called gates. A LSTM cell is shown in Figure 2. It consists of an input gate (ig), a forget gate (fg), an output gate (og) and a memory cell (mc) (shown in Equation 3). These gates can learn to modulate the flow of data, depending on the input and the hidden states, which helps in imitating long-range dependencies.
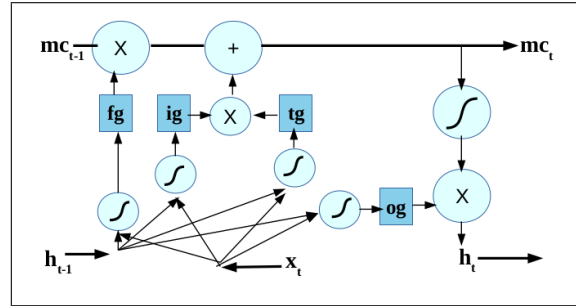

Fig. 2: A long short term memory cell

$$\text{ig}_t = sigmoid(\text{RecWeight}_{\text{ig}}\text{h}_{t-1} + \text{Weight}_{\text{ig}}\text{x}_t + \text{bias}_{\text{ig}})$$
$$\text{fg}_t = sigmoid(\text{RecWeight}_{\text{fg}}\text{h}_{t-1} + \text{Weight}_{\text{fg}}\text{x}_t + \text{bias}_{\text{fg}})$$
$$\text{og}_t = sigmoid(\text{RecWeight}_{\text{og}}\text{h}_{t-1} + \text{Weight}_{\text{og}}\text{x}_t + \text{bias}_{\text{og}})$$
$$\text{tg}_t = hypTan(\text{RecWeight}_{\text{mc}}\text{h}_{t-1} + \text{Weight}_{\text{mc}}\text{x}_t + \text{bias}_{\text{mc}})$$
$$\text{mc}_t = \text{fg}_t \circ \text{mc}_{t-1} + \text{ig}_t \circ \text{tg}_t$$
$$\text{hidden}_t = \text{og}_t \circ hypTan(\text{mc}_t), \tag{3}$$

where fg_t, ig_t, og_t, mc_t are the activation at the forget gate, input gate, output gate, and memory cell at time instant t. RecWeight_r and Weight_r are the learned weight of

recurrent and input connections and bias$_r$ are the respective biases. The subscript r may be any of the forget, input, output, or memory cell. The activation function used are sigmoid and hyperbolic tangent (hypTan). Finally $\circ$ symbolizes for element-wise multiplication of the matrices, and hidden$_t$ represents the hidden state vector or the output of the LSTM unit at time instant t.

The proposed LSTM classification model consists of two LSTM layers, followed by a fully connected layer with soft-max activation for classifying the daily rainfall.

(a) Input layer- The input to the first LSTM layer is of the form ($BatchSize \times TimeStepHist \times NumFeatures$).
(b) Hidden layer- There are two hidden layers with 80 neurons in each LSTM cell. The activation function used is hyperbolic tangent.
(c) Dense softmax layer- A fully connected layer connecting the last output of LSTM with softmax activation to classify monsoon days into classes (wet, normal, or dry).

The hyper-parameters $BatchSize$ and $TimeStepHist$ are fixed empirically, by varying different length of history at lead over the validation period.

Sequence-to-sequence model: Sequence to Sequence (Seq2Seq) is a composite model used to learn time-series dependencies. The model is built with two LSTM units, one imitating the functioning of an encoder, while the other imitate a decoder. As usual, the encoder tries to learn from the input sequence, and it produces a complex hidden learned state or context, which aims to apprehend the variability of the input. The decoder tries to produce the sequence of output, considering the learned complex context by the encoder and the previous output state. Attention mechanism as proposed by Bahdanau et al. [2] adds more power to the present sequence-to-sequence model. During decoding, the attention mechanism permits the decoder to concentrate over different encoder outputs with addition to the last encoder state. A set of attention weights is calculated, which is then multiplied with the encoder outputs to create context vectors. The context vectors contain information pertaining to a certain part of the sequence in input, and thus helping the decoder in producing a more accurate output.

The proposed Seq2Seq monsoon classification model, consists of input similar to LSTM (discussed previously), two LSTM units with 100 neurons each, in encoder and decoder parts of the model and finally, a dense soft-max layer. The attention mechanism is also applied to improve the model's performance. The proposed architecture and working of Seq2Seq model is shown in Figure 3.

The training of Seq2Seq classification model is performed as following. The input sequence is fed to the first encoder layer. The last cell of the encoder stores the contextual information related to the sequence, which is fed to the decoder. During decoding steps, the decoder output of the previous time-step $t-1$ is provided as an input at the succeeding time-step t. This method is known as *greedy sampling* and makes the model more robust to learn from previous mistakes. An alternative to this procedure is to feed the correct input to the decoder at every time-step, even if the decoder have made a mistake at previous step. This method is known as *guided training*. We have implemented both the training methods for Seq2Seq model, however, *greedy sampling* performed better for the rainfall classification task.
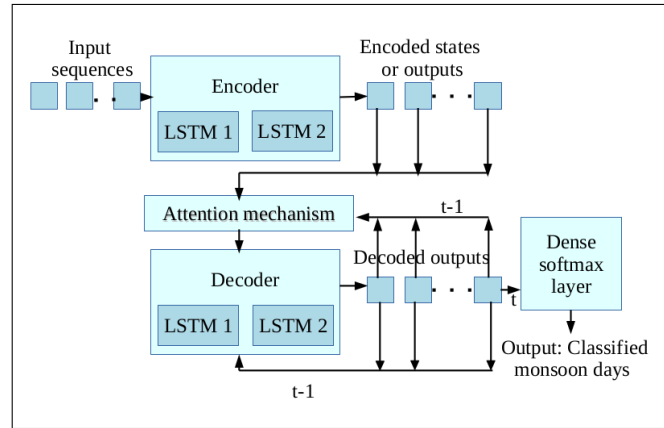
Fig. 3: Sequence-to-sequence model with attention mechanism for daily rainfall classification

Experimental setup- sliding window: Real time classification of dry, wet, or normal days requires continuous learning of sequential input. The number of time-steps to be learned increases with time and the computational complexity becomes exponential. Additionally, the near time-steps influence the current output value to a greater extent as compared to the time-steps at far past. For these reasons, a sliding window method is used (Figure 4) in training process of the proposed approach. A sliding window of size $TimeStepsHistory$ is chosen, which indicates a fixed number of previous time-steps at lead is used to predict the following time-step.
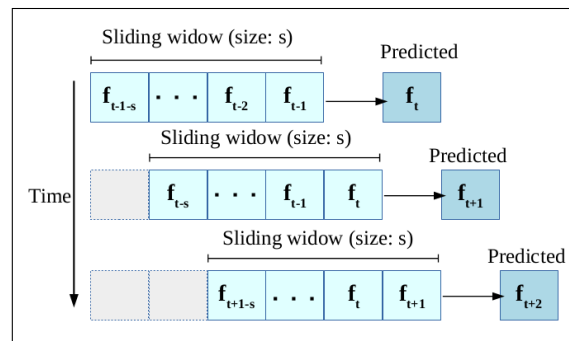

Fig. 4: Sliding window training approach (eg. prediction of f_t instant is provided using past time-steps from f_{t-1} to f_{t-1-s}, i.e. TimeStepsHistory = s time-steps (sliding window size))

Experimental setup- hyper parameters: For LSTM and Seq2Seq models, the optimal values for various hyper-parameters are ascertained empirically. The values are selected considering their performance on the validation data set. Hyper-parameters for the model include the count of layers, the count of neurons in LSTM cell, the length of input sequence history ($TimeStepsHist$) or sliding window size to be considered, and weighting component. The learning rate value is set to $1 * e^{-3}$ for training. Mini-batch is considered with size of 32 for gradient descent. Adam optimizer is used, and finally, L2 regularization is implemented with the regularization parameter equal to $1 * e^{-3}$ to avoid over-fitting of the model.

Experimental setup- loss function: A soft-max cross entropy loss is used for modeling the multi-class classification problem. However, there is a problem of class imbalance for our study. Out of the total days considered, 15% are wet days, 18% are dry days and rest 67% resembles the normal days. To solve the issue of class imbalance , weighted soft-max cross entropy loss is used, which allows to use different scores for different classes, and hence a mis-classification of the less common classes can be penalized more to improve prediction accuracy. For Seq2Seq model, the cost function is considered as a amalgamation of two losses– (a) loss 1: average loss of each time step prediction, (b) loss 2: loss of the prediction calculated at the last time step. The losses are combined as shown in Equation 4, where β represents the weighting factor.

$$\text{overall loss} = \beta * \text{loss } 1 + (1\text{-}\beta) * \text{loss } 2 \qquad (4)$$

4   Experimental Results and Analysis

The classification of daily rainfall into dry, wet or normal days are presented with measures, namely, precision, recall, accuracy, and f1-score. The outcome of break and active monsoon spells detection are also discussed in the following section.

4.1   Training and Test Sets

Data are divided into three groups, namely, training group, validation group, and test group of samples. The proposed models are designed and optimized utilizing the data in the training group. Values for different hyper-parameters are selected based on their performance of outcomes over the data samples of the validation group, and finally, the model is assessed using the samples in the test group. The input consists of 8174 time-steps, among which 80% is used for training the models, 10% for setting the hyper-parameters (validation) and the remaining 10% for test and analysis. In terms of years, approximately, fifty-three years of sixty-seven are consumed for training, seven for validation and seven in testing. As required by the problem, only days of June-September of all the years represent the tuples. Among 8174 days under consideration, the number of wet and dry days are observed as 1211 and 1470, which comprised 14.8% and 18% of total number of days under study.

4.2   Performance Metrics

The multi-class monsoon classification problem is evaluated with precision and recall of each class (dry, wet or normal). Precision is computed as the count of accurately classified positive class among all the results that it has classified as positive. Similarly, recall is computed as the count of rightly classified positive class out of observed or actual positive class examples. The overall classification result is presented in terms of accuracy and f1-score, where each of these measures are calculated as average of same measures for individual classes. F1-score refers to the harmonic mean of recall and precision. F1-score is better as its value approaches 1 and worst at value 0.

4.3   Classification Results of Daily Monsoon

The performances of the LSTM and the Seq2Seq model in classifying the daily rainfall classes is presented in terms of performance metrics (discussed in previous section) for the test period 2008-2014. The classification is performed considering a lead of one to five days. Conventional classifiers, namely, the support vector machine (SVM) and K-nearest neighbor (KNN) are also designed for evaluating the proposed models against them. For the SVM classifier model, various kernels are implemented but the linear kernel is observed to provide best results. For the KNN model, the number of neighbors is considered as a hyper-parameter and varied with lead times. The classification results by the proposed models along with the two traditional models are presented for a lead of one to five days in Tables 1(a-e), respectively. It is observed that classification performance is best for least lead of one day and it degrades gradually as the lead is increased from one to five days. Dry days are noted to be classified with better accuracy as compared to the wet days. The proposed LSTM classifier provides an accuracy of 92.73% and f1-score of 0.953 at a lead of one day. Similarly, the Seq2Seq classifier classify monsoon days with 92.50% accuracy and 0.953 f1-score. Moreover, it classify the dry and wet days with precision of 0.952 and 0.856, respectively. The proposed Seq2Seq model classify the daily rainfall with best accuracy among all the four models. SVM classifier with linear kernel performs comparable, but the KNN classifier can not classify the monsoon days with good accuracy. The proposed models are observed to perform superior to the SVM and KNN classifiers for all the leads. The variation of accuracy and f1-score of the classification by proposed LSTM and Seq2Seq models, and traditional SVM and KNN classification models are shown in Figures 5a and 5b, respectively.

4.4   Break and Active Spells Detection

The classification models predict the occurrence of dry or wet days, which is extended to the prediction of break or active spell. The classified days are summed up to present the break or active spells. We summarize a break or an active spell only when we obtain continuous three or more days (as per their definition) of dry or wet classified days, respectively. There are sixteen break and twenty-one active spells in the test period (2008-2014).

Table 1: Precision and recall for classifying dry, wet, or normal days and overall accuracy and f1-score of classification at different leads (a-e: prediction at a lead of one to five days) for 2008-2014

(a) Classification at lead 1

Models	SVM	KNN	LSTM	Seq2Seq
Dry day classification				
Precision	0.931	0.810	0.951	0.952
Recall	0.955	0.942	0.950	0.982
Wet day classification				
Precision	0.841	0.742	0.870	0.856
Recall	0.835	0.751	0.884	0.884
Normal day classification				
Precision	0.944	0.923	0.961	0.967
Recall	0.947	0.877	0.962	0.953
Overall classification				
Accuracy	90.53	82.50	92.73	92.50
F1-score	0.935	0.875	0.954	0.953

(b) Classification at lead 2

Models	SVM	KNN	LSTM	Seq2Seq
Dry day classification				
Precision	0.869	0.756	0.885	0.921
Recall	0.936	0.770	0.936	0.961
Wet day classification				
Precision	0.807	0.651	0.803	0.851
Recall	0.752	0.462	0.842	0.801
Normal day classification				
Precision	0.925	0.820	0.944	0.943
Recall	0.921	0.87	0.922	0.936
Overall classification				
Accuracy	86.70	74.23	87.73	90.50
F1-score	0.903	0.795	0.910	0.926

(c) Classification at lead 3

Models	SVM	KNN	LSTM	Seq2Seq
Dry day classification				
Precision	0.793	0.633	0.809	0.909
Recall	0.808	0.596	0.840	0.834
Wet day classification				
Precision	0.784	0.319	0.782	0.796
Recall	0.672	0.252	0.652	0.744
Normal day classification				
Precision	0.871	0.741	0.882	0.911
Recall	0.897	0.781	0.901	0.932
Overall classification				
Accuracy	81.60	56.43	82.43	87.20
F1-score	0.849	0.667	0.859	0.890

(d) Classification at lead 4

Models	SVM	KNN	LSTM	Seq2Seq
Dry day classification				
Precision	0.696	0.541	0.670	0.751
Recall	0.701	0.509	0.670	0.808
Wet day classification				
Precision	0.536	0.219	0.614	0.686
Recall	0.363	0.192	0.421	0.578
Normal day classification				
Precision	0.784	0.721	0.823	0.836
Recall	0.832	0.742	0.855	0.862
Overall classification				
Accuracy	67.20	49.36	70.23	75.76
F1-score	0.745	0.615	0.769	0.814

(e) Classification at lead 5

Models	SVM	KNN	LSTM	Seq2Seq
Dry day classification				
Precision	0.641	0.434	0.681	0.744
Recall	0.605	0.318	0.601	0.707
Wet day classification				
Precision	0.333	0.301	0.525	0.531
Recall	0.272	0.231	0.34	0.413
Normal day classification				
Precision	0.741	0.684	0.761	0.823
Recall	0.797	0.765	0.845	0.842
Overall classification				
Accuracy	57.16	47.30	65.56	69.93
F1-score	0.684	0.603	0.717	0.759

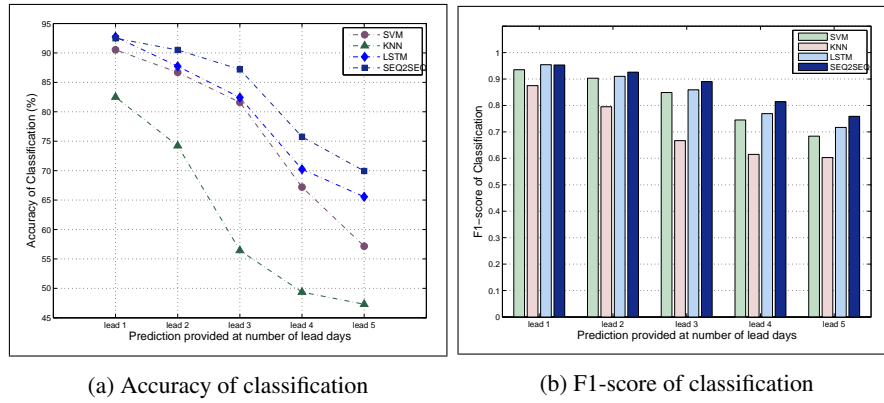(a) Accuracy of classification   (b) F1-score of classification

Fig. 5: Comparison of (a) accuracy and (b) f1-score in monsoon classification by proposed LSTM and Seq2Seq, with conventional SVM and KNN classifiers at lead of one to five days for 2008-2014

Table 2 highlights the total count of observed spells and count of correctly identified spells by LSTM and Seq2Seq models at lead of three days. It is noted that LSTM and Seq2Seq models have correctly identified thirteen and sixteen out of twenty-one active spells, respectively. Similarly, the corresponding models have correctly identified thirteen and fourteen break spells, respectively. Identification of break spells appear to be predominant to the identification of active spells (similar to the trend of dry and wet days classification). A detailed evaluation is presented elaborating the actual length of each spells (as number of days) along with the continuous days predicted as respective classes (wet or dry) by the proposed models. The observed length of spells and corresponding predicted period of days for break and active spells during 2008-2014 are shown in Tables 3a and 3b, respectively. It is noted that the Seq2Seq model predict the dry and wet spells more accurately in terms of length of spells. Although the LSTM model detects dry spells quite well, but it fails to detect many of the active spells.

Table 2: Observed and correctly predicted count of break and active monsoon spells at a lead of three days by LSTM and Seq2Seq models for test-period 2008-2014

Models	Observed count of break spells	Predicted count of break spells	Observed count of active spells	Predicted count of active spells
LSTM	16	13	21	13
Seq2Seq	16	14	21	16

The observed length and the predicted length of break and active monsoon spells are also presented in Figures 6a and 6b, respectively. Actual length of observed break or active spells in terms of number of continuous days is shown by the bars in the figure, and the symbols represent the predicted spell length by the proposed LSTM and Seq2Seq models.

Table 3: Length of observed and predicted break (BS) and active (AS) monsoon spells (as number of days) at lead of three days by LSTM and Seq2Seq models for test-period 2008-2014

(b) Active spells

	No. of days in observed active spell	No. of days predicted by LSTM	No. of days predicted by Seq2Seq
AS 1	3	2	3
AS 2	3	2	2
AS 3	3	3	3
AS 4	4	4	4
AS 5	3	3	3
AS 6	4	1	2
AS 7	4	1	3
AS 8	8	5	6
AS 9	6	3	3
AS 10	3	0	2
AS 11	6	3	5
AS 12	4	3	3
AS 13	4	3	4
AS 14	4	4	4
AS 15	3	2	2
AS 16	11	7	9
AS 17	5	3	4
AS 18	6	5	5
AS 19	3	1	3
AS 20	4	3	4
AS 21	4	2	2

(a) Break spells

	No. of days in observed break spell	No. of days predicted by LSTM	No. of days predicted by Seq2Seq
BS 1	30	28	28
BS 2	10	6	8
BS 3	13	13	13
BS 4	7	7	6
BS 5	5	3	4
BS 6	14	13	13
BS 7	16	15	15
BS 8	6	5	5
BS 9	9	7	7
BS 10	9	6	4
BS 11	15	11	12
BS 12	4	2	3
BS 13	3	0	0
BS 14	5	3	4
BS 15	3	0	2
BS 16	6	4	5

5   Conclusions

The prediction of break and active monsoon spells of India is performed during June-September. Two different models , namely the LSTM classification model and the Seq2Seq model with attention mechanism are proposed. Both the models performed better than traditional SVM and KNN classifiers. One of the important observations is that detecting active spells is harder than detecting dry spells. The possible explanation can be that an average dry spell lasted longer compared to an average active spell, thus making it less random and easier to detect. Another major problem faced during the classification process is that the data set is imbalanced. Weighted soft-max loss is implemented but the weights add to number of hyper-parameters thus making the model harder to tune. Even though the hyper-parameters are selected considering performance over the validation set, grid search or other similar methods can be implemented to detect the optimal set of hyper-parameters. LSTM classification model and the Seq2Seq model can also be greatly improved by acquiring more data. Another possible area of
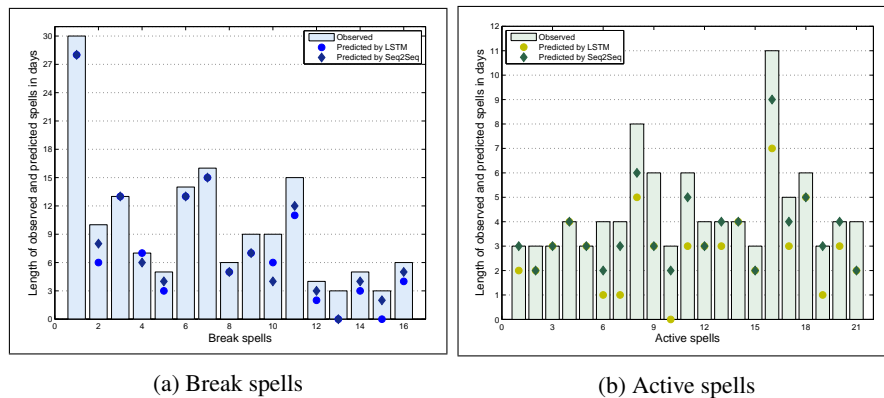
(a) Break spells

(b) Active spells

Fig. 6: Observed and predicted length of monsoon spells by LSTM and Seq2Seq models during test period 2008-2014 for (a) break spells, and (b) active spells

extension can be use of convolution neural network (CNN) for multivariate time series. Although CNNs have extensively used as feature extractors in images, they can also be used to extract features from climatic time-series. Hybrid of CNN and LSTM can also be used for further prospect. Finally, exploration of spatial-temporal nature of climate data may lead to better understanding, and classifying or predicting the climatic phenomenon.

References

1. Abhilash, S., Sahai, A., Pattnaik, S., Goswami, B., Kumar, A.: Extended range prediction of active-break spells of Indian summer monsoon rainfall using an ensemble prediction system in ncep climate forecast system. International Journal of Climatology **34**(1), 98–113 (2014)
2. Bahdanau, D., Cho, k., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations. p. http://arxiv.org/abs/1409.0473 (2015)
3. Chattopadhyay, S., Chattopadhyay, G.: Comparative study among different neural net learning algorithms applied to rainfall time series. Meteorological Applications **15**(2), 273–280 (2008)
4. Fan, K., Liu, Y., Chen, H.: Improving the prediction of the East Asian Summer Monsoon: New approaches. Weather & Forecasting **27**(4), 1017–1030 (2012)
5. Gadgil, S., Joseph, P.V.: On breaks of the Indian monsoon. Journal of Earth System Science **112**(4), 529–558 (2003)
6. Gadgil, S., Rajeevan, M., Nanjundiah, R.S.: Monsoon prediction- why yet another failure? Current Science **88**(9), 1389–1400 (2005)
7. Hong, W.C.: Rainfall forecasting by technological machine learning models. Applied Mathematics and Computation **200**(1), 41–57 (2008)
8. Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R., Joseph, D.: The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society **77**(3), 437–471 (1996)

9. Kiran, V.R., Rajeevan, M., Rao, S.V.B., Rao, N.P.: Analysis of variations of cloud and aerosol properties associated with active and break spells of Indian summer monsoon using MODIS data. Geophysical Research Letters **36**(9) (2009)

10. Kripalani, R.H., Kulkarni, A.: Rainfall variability over South–East Asia–connections with Indian monsoon and ENSO extremes: new perspectives. International Journal of Climatology **17**(11), 1155–1168 (1997)

11. Kucharski, F., Bracco, A., Yoo, J.H., Molteni, F.: Atlantic forced component of the Indian monsoon interannual variability. Geophysical Research Letters **35**(4) (2008)

12. Kumar, K.K., Rajagopalan, B., Hoerling, M., Bates, G., Cane, M.: Unraveling the mystery of Indian monsoon failure during El Niño. Science **314**(5796), 115–119 (2006)

13. Kumar, M.R.R., Dessai, U.R.P.: A new criterion for identifying breaks in monsoon conditions over the Indian subcontinent. Geophysical research letters **31**(18), L18201 (2004)

14. Mandke, S.K., Sahai, A.K., Shinde, M.A., Joseph, S., Chattopadhyay, R.: Simulated changes in active/break spells during the Indian summer monsoon due to enhanced CO2 concentrations: assessment from selected coupled atmosphere–ocean global climate models. International journal of climatology **27**(7), 837–859 (2007)

15. Mekanik, F., Imteaz, M.A., Gato-Trinidad, S., Elmahdi, A.: Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. Journal of Hydrology **503**, 11–21 (2013)

16. Prasad, V.S., Hayashi, T.: Active, weak and break spells in the Indian summer monsoon. Meteorology and Atmospheric Physics **95**(1-2), 53–61 (2007)

17. Rajeevan, M., Bhate, J., Kale, J.D., Lal, B.: High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells. Current Science **91**(3), 296–306 (2006)

18. Rajeevan, M., Gadgil, S., Bhate, J.: Active and break spells of the Indian summer monsoon. Journal of Earth System Science **119**(3), 229–247 (2010)

19. Ramamurthy, K.: Monsoon of India: some aspects of the break in the Indian southwest monsoon during July and August. Forecasting manual **157** (1969)

20. Rao, T.N., Saikranthi, K., Radhakrishna, B., Rao, S.V.B.: Differences in the climatological characteristics of precipitation between active and break spells of the Indian summer monsoon. Journal of Climate **29**(21), 7797–7814 (2016)

21. Saha, M., Chakraborty, A., Mitra, P.: Predictor-year subspace clustering based ensemble prediction of Indian summer monsoon. Advance in Meteorology **2016**(9031625), 1–12 (2016)

22. Saha, M., Mitra, P.: Identification of Indian monsoon predictors using climate network and density-based spatial clustering. Meteorology and Atmospheric Physics **128**(5), 1–14 (2018)

23. Saha, M., Mitra, P., Chakraborty, A.: Fuzzy clustering-based ensemble approach to predicting Indian monsoon. Advance in Meteorology **2015**(32835), 1–12 (2015)

24. Saha, M., Mitra, P., Nanjundiah, R.S.: Autoencoder-based identification of predictors of Indian monsoon. Meteorology and Atmospheric Physics **https://doi.org/10.1007/s00703-018-0637-y**, 613–628 (2016)

25. Saha, M., Mitra, P., Nanjundiah, R.S.: Prediction of active and break spells of Indian summer monsoon using linear discriminant analysis. In: International Conference on Advances in Pattern Recognition. pp. 1–6 (2017)

26. Saha, M., Mitra, P., Nanjundiah, R.: Predictor discovery for early-late Indian summer monsoon using stacked autoencoder. Procedia Computer Science **80**, 565–576 (2016)

27. Saha, M., Mitra, P., Nanjundiah, R.: Deep learning for predicting the monsoon over the homogeneous regions of India. Journal of Earth System Science **126**(4), 1–18 (2017)

28. Sudeepkumar, B.L., Babu, C.A., Varikoden, H.: Future projections of active-break spells of Indian summer monsoon in a climate change perspective. Global and Planetary Change **161**, 222 – 230 (2018)