A. SUMMARY SHEET

1. TITLE OF THE PROJECT:

**Forecasting of Rainfall in India Using Large-Scale Climate Indices through Advanced Deep Learning Approach**

2. REFERENCE NO. OF SANCTION LETTER WITH DATE: SRG/2020/001871

3. PI NAME & ORGANISATION:

(COMPLETE ADDRESS WITH TELEPHONE NO. FAX & EMAIL

DETAILS)

**Dr. Vinit Jakhetiya**

**Assistant Professor, Department of CSE,**

**Indian Institute of Technology Jammu, J&K, India.**

**Mob: 97973886090**

**Email:** [**vinit.jakhetiya@iitjammu.ac.in**](mailto:vinit.jakhetiya@iitjammu.ac.in)

4. DATE OF START:

5. DATE OF COMPLETION: SCHEDULED AND LIKELY

6. TOTAL COST OF THE PROJECT: SANCTIONED AND NOW ESTIMATED

7. STAFF SANCTIONED & IN POSITION:

**Sanctioned:**

**In Position:**

1. **Junior Research Fellow**

8. TOTAL EXPENDITURE:

9. ASSETS DEVELOPED OR EQUIPMENTS ACQUIRED, IF ANY:

10. SUMMARY OF PROGRESS AGAINST APPROVED WORK-PLAN/TIME

SCHEDULE OF ACTIVITIES IN THE PROJECT:

Approved Work Plan:

Summary of Progress:

11. ISSUES NEEDING ATTENTION OF GOVERNMENT/LOCAL BODIES:

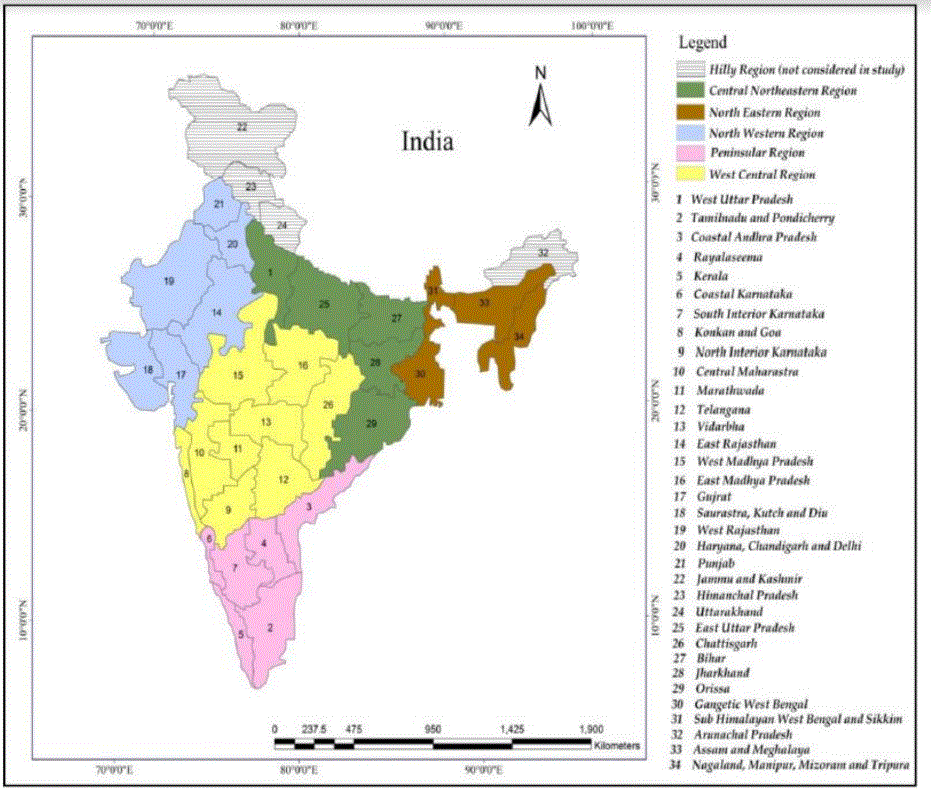
N/A

B. DETAILED REPORT

(For the period from to )

1. INTRODUCTION (NEED ASSESSMENT FOR S&T INTERVENTION IN PROJECT AREA):

In India, agriculture still forms a large portion of the economy, which is directly dependent on rainfall. Precipitation is mostly consistent over the nation during the southwest monsoon. Approximately 88 cm of rain falls on average each season (1961-2010 average value), with a variance coefficient of about 10%. (Drought) years are years with rainfall above (less than) 110% (90%) of average. Presently, IMD (Indian Meteorological Department) is studying the relationships between potential climate parameters and ISMR and various machine learning models using EMR (Ensemble Multiple Regression) and PPR (Projection Pursuit Regression). ISMR is forecasted using a sliding window, as well as ML models using 10 climate parameters.

**Fig 1**. 36 sub-divisions of rainfall forecasting in whole India.

A large number of intricate features of data were acquired through the use of an auto-encoder, which was used to reduce the dimensionality of the data. ANN model is used for recreating the input parameter and for obtaining the 6 major correlated vectors that are required to train the ML model with the 0-40 years sliding window. A correlation coefficient of 0.829 was found between the ISMR data and the rainfall model forecast.

The difference between the seasonal mean rainfall and ISMR data for the period (1991-2019) reveals recent deficiency, normal, and excess monsoon years. Based on 36 meteorological subdivisions, 7 homogeneous regions and all of India as a whole, monthly rainfall data is analyzed for historical Indian rainfall.

2. APPROVED OBJECTIVES OF THE PROJECT:

1. Exploit the growth of latest Machine/Deep Learning techniques such as bi-directioanl LSTM, GRU, Deep Echo State Network (DeepESN) and fusing these techniques via ensemble learning.
2. Predict the rainfall in each sub-region in India, for analyzing the drought and flood in the particular part of the country.
3. Analyze forecasting of the rainfall in all subdivisions and classifying them into five homogenous regions of normal rainfall, excessive rainfall, flood, drought, less rainfall.

3. PROJECT AREA (BLOCK, VILLAGE, TOTAL AREA COVERED)

4. COMMUNITY BACKGROUND AND KNOWLEDGE LEVEL (CASTE, OCCUPATION,

INCLUDING TRADITIONAL KNOWLEDGE & PRACTICES FOLLOWED)

**N/A**

5. METHODOLOGY & SYSTEMS APPROACH (SURVEY/ PRA EXERCISE; COMMUNITY MOBILIZATION & SOCIAL ENGINEERING; TECHNOLOGY IDENTIFICATION, MODULATION & DIFFUSION & TRAINING COMPONENT, ETC.):

**Objective 1:** Exploit the growth of latest Machine/Deep Learning techniques such as bi-directioanl LSTM, GRU, Deep Echo State Network (DeepESN) and fusing these techniques via ensemble learning.

**Data Gathering And Preprocessing**

The climate variables examined for the proposed approach of identifying new predictors of monsoon, and thereby predicting the Indian summer monsoon are Sea Surface Temperature (SST), Sea Level Pressure (SLP), and Zonal u-wind (UWND) at 200hPa pressure level.

The datasets were taken from [here](https://iridl.ldeo.columbia.edu/). These datasets were in NetCDF(\*.nc) format. SLP and UWND data were provided at 2.5° × 2.5° grid location. SST data were provided in 2° × 2° resolution. Data from each month from 1948-2020 are considered. Using simple python scripts, NetCDF files were converted to .csv files and the data was re-grided to 10° × 20° grid locations which adds up to 324 rectangular grid boxes for SLP and UWND and 192 for SST.

Then we segregated the data into training and testing sets. The training set is utilized to discover the monsoon predictors using stacked autoencoders and develop a prediction model for the monsoon. The test data set is used to evaluate the performance of new predictors. The training period is 1948-2000 and the test period is 2001-2020.

The monthly anomalies(denoted as Anomaly my for the month m and year y) were calculated, by the following equation, considering the years 1961-1990 as base years.

*Anomaly my = X my - mean(Xm),*

where X my is the variable for the mth month and yth year and the mean(Xm) denotes the average of the mth month over base period.

**Objective 3:**

**(a) Development of Deep learning model**

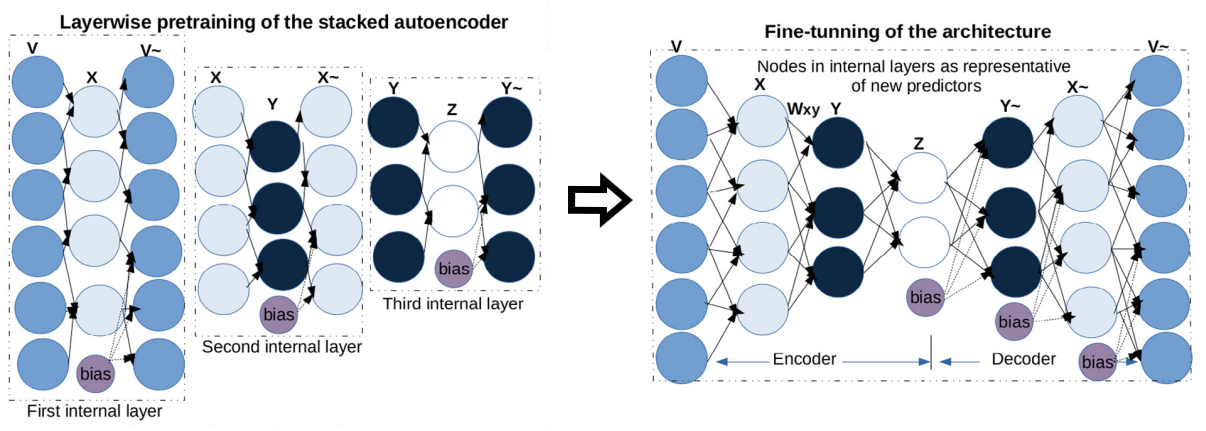
Instead of taking a single‐model output, an amalgam of the multiple neural networks is being designed for the rainfall forecasting problem. We hypothesize, It can get the best features from the models and then assemble them into a single recursive network. Such techniques are being utilized for improving the accuracy, better generalization and providing optimized methods for removal of uncertainties. In order to get better results, we are using different combinations of loss functions, optimizers and activations.

Basic backbone models are briefly described:

1. **Stacked Autoencoders**

An autoencoder is an architecture belonging to the class of artificial neural networks. It has an input, an internal layer, and an output layer. We used this architecture to acquire the intricate features of data, and utilized it for the reduction in dimensionality of the data.The autoencoder sets the output values to be the same as the input values, and attempts to recreate the input information from the representation learned in the internal layer. The architecture is capable of learning a non-linear function by using the process of iterative training of the model. The model aims at reducing the re-construction errors in rebuilding the output from the representation discovered in the internal layer.

Several single-layered autoencoders are heaped to generate a deep neural model. The outcome of the first autoencoder acts as an input to the second, and this continues to the desired depth of the architecture. We show the structure of the stacked autoencoder and its training method in Fig 1



**Fig 2**. Three autoencoders are pre-trained and then stacked to form the full architecture.

The training of the stacked autoencoder is unsupervised, which does not require any information about the predictand variable. The pre-training is performed, taking into consideration a single layer at a time. The layers of the stacked autoencoder are trained with the motivation of minimization of errors in input reconstruction individually for the layers (shown in the left portion of Fig 2).

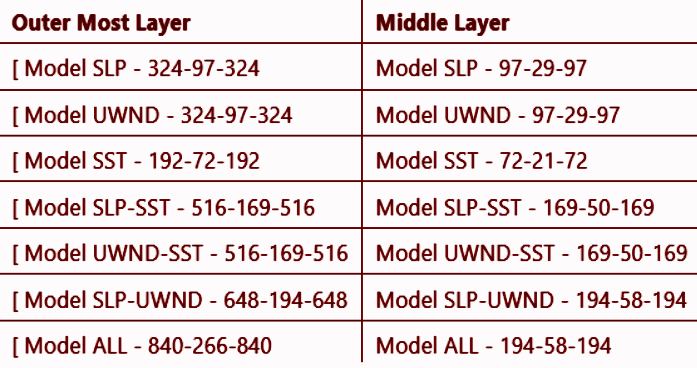
After pre-training all the layers, the whole stacked network is fine-tuned with a run of gradient descent, and the weights assigned to each layer are adjusted (shown in the right portion of fig 1). We learned novel complex features in the internal layers of the deep autoencoder model.

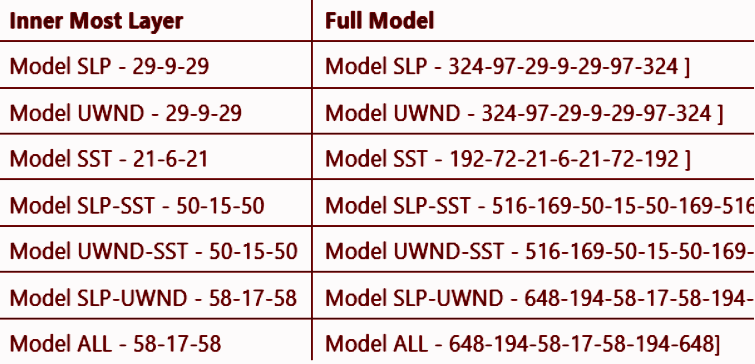
**(b) Identifying The Predictors**

The predictor identification process initiates with unsupervised training and fine-tuning of a stacked autoencoder, followed by a threshold over the acquired weights. Finally, the required predictors are filtered from their correlation study with the Indian summer monsoon.

We identified two types of predictors using a stacked autoencoder, the first from individual climatic variables and the second from a combination of variables.

The model built for 7 different combinations of input variables are stored in a 2-D list and used whenever required. The architecture along with the input variable combination is shown in the following figure. The outermost layer is the outer autoencoder, middle layer is the intermediate autoencoder and the innermost layer is the inner autoencoder as shown in Fig 3.





**Fig 3.** In every row three columns define three autoencoders with a mentioned number of nodes which are later stacked to form the autoencoder mentioned in the last column.

1. *Training of stacked autoencoders for monsoon predictor identification*

We examined input data for predictor identification from 1948–2000 for developing the stacked autoencoder architecture. The input corresponds to the data at a monthly scale, which adds to 636 (53 years × 12 months) instances. Unsupervised pre-training of the three autoencoders is performed taking output of the middle layer of an autoencoder as input for the next autoencoder. Then fine tuning of the stacked model is done after transferring weights of the pretrained autoencoders to the stacked model. We trained the layers by reducing the error in reconstructing the input following the principle of the autoencoder. The nodes learned in the internal levels of the stacked autoencoder present the composite and sophisticated features.

1. *Post-treatment by thresholding of weights*

After fine tuning, from the internal layers of the stacked model we acquire three sets of predictors. The threshold is adapted to consider the input nodes that actively influence the node in the inner layer while discarding the rest. We put a limit on the weight matrix by considering the weights with a value higher than twice the standard deviation over the mean learned in the weight matrix. This contributes to an evaluation of the node in the internal layer, which is the seed for a new potential monsoon predictor. New climatic predictors are evaluated as the weighted sum of the nodes in the input layer that are being selected after the threshold method.

1. *Post-filtering of identified monsoon predictors*

The newly-identified predictors of each layer are ranked by studying the predictors’ (corresponding to the nodes of the internal layers) correlation with the Indian summer monsoon. This considers a lead of one to twelve months for finding the highest correlated month for all the identified predictors using the Pearson correlation coefficient (γ). We found out that May is highly correlated to the identified predictors. We considered the best lead month(May) of the identified predictors for further evaluation.

**Experimental Results:**

| **Sr** | **Model**  **SLP** | | **Model UWND** | | **Model**  **SST** | | **Model**  **SLP-SST** | | **Model**  **UWND-SST** | | **Model**  **SLP-UWND** | | **Model**  **ALL** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | -0.6685 | -0.0744 | -0.6531 | 0.3051 | -0.5375 | -0.3075 | -0.6413 | 0.2675 | -0.6506 | 0.1478 | -0.6423 | 0.0746 | -0.6942 | 0.1765 |
| **2** | -0.6679 | 0.3934 | -0.5791 | 0.2030 | -0.5374 | -0.3099 | -0.6733 | 0.1777 | -0.6265 | -0.2565 | -0.5473 | 0.4506 | -0.7840 | 0.1252 |
| **3** | -0.6295 | 0.1169 | -0.5375 | 0.2438 | -0.5378 | -0.2911 | -0.5828 | 0.2916 | -0.6339 | 0.1102 | -0.7482 | 0.3208 | -0.5554 | -0.0032 |
| **4** | -0.6409 | 0.2043 | -0.6104 | 0.3537 | -0.5375 | -0.3092 | -0.5495 | 0.1829 | -0.5711 | 0.0343 | -0.6554 | 0.0134 | -0.6905 | -0.0079 |
| **5** | -0.6338 | -0.0090 | -0.7763 | -0.0714 | -0.5375 | -0.3102 | -0.6416 | 0.5348 | -0.6534 | 0.0023 | -0.5472 | -0.0764 | -0.6253 | -0.0740 |
| **6** | -0.7104 | 0.4662 | -0.7004 | 0.2021 | -0.5375 | -0.3004 | -0.6865 | 0.2030 | -0.6254 | -0.0972 | -0.6431 | -0.1050 | -0.6451 | -0.0202 |
| **7** | -0.5543 | 0.5390 | -0.6875 | 0.2825 | -0.5394 | -0.3096 | -0.5200 | 0.3109 | -0.5852 | 0.0050 | -0.5822 | -0.0634 | -0.5696 | 0.3826 |
| **8** | -0.5864 | 0.1666 | -0.6369 | 0.1401 | -0.5375 | -0.3094 | -0.5896 | 0.3871 | -0.6283 | 0.0239 | -0.5660 | 0.2438 | -0.5887 | 0.1896 |
| **9** | -0.7181 | 0.2057 | -0.7073 | 0.2052 | -0.5375 | -0.3091 | -0.5375 | 0.2509 | -0.6149 | 0.1392 | -0.6105 | 0.1181 | -0.6130 | 0.4280 |
| **10** | **-0.8428** | 0.2959 | -0.6980 | 0.4741 | -0.5375 | -0.3105 | -0.6616 | 0.1312 | -0.7391 | -0.3449 | -0.6932 | 0.3319 | -0.5765 | -0.0441 |
| **11** | -0.7855 | 0.3144 | -0.6090 | 0.6531 | -0.5376 | -0.3093 | -0.5375 | 0.1901 | -0.6957 | 0.0236 | -0.6176 | 0.3538 | -0.6907 | -0.0275 |
| **12** | -0.5157 | 0.4167 | -0.5983 | 0.3052 | -0.5375 | -0.3103 | -0.6532 | 0.1881 | -0.6489 | 0.4909 | -0.7548 | -0.1224 | -0.5515 | 0.0254 |
| **13** | -0.7120 | 0.4824 | **-0.8585** | 0.3979 | -0.5382 | -0.3105 | -0.5375 | 0.3340 | -0.7664 | -0.0394 | **-0.8110** | 0.4267 | -0.6887 | -0.0947 |
| **14** | -0.7054 | 0.3370 | -0.6339 | 0.0247 | -0.5381 | -0.3105 | -0.6167 | 0.2695 | -0.7282 | 0.1389 | -0.6672 | 0.2077 | -0.6337 | -0.0379 |

6. TECHNICAL BACK-UP SUPPORT (NAMES OF SCIENTISTS INVOLVED AND SUPPORT RECEIVED & LINKAGES ESTABLISHED WITH S&T INSTITUTIONS):

**N/A**

7. SCIENCE & TECHNOLOGY COMPONENT (TECHNOLOGY PACKAGE DEVELOPMENT/ NEW INNOVATIONS/ OBSERVATIONS):

**We observed high correlations between some of the obtained predictors and rainfall. The predictors at deep layers outperform predictors obtained in the shallow layer. Higher predictability of predictors from a deeper layer highlights the advantages of using a stacked autoencoder over a single layer autoencoder.**

8. PEOPLE’S PARTICIPATION FROM PLANNING TO IMPLEMENTATION STAGE (WITH EMPHASIS ON THEIR INVOLVEMENT IN TECHNOLOGY GENERATION/ MODULATION/ TRANSFER/ ADOPTION; CO-OPERATIVE FORMATIONS/ SELF HELP GROUPS: GENDER PERSPECTIVE):

**N/A**

9. INDICATORS APPLIED FOR MONITORING (QUALITATIVE & QUANTITATIVE

ANALYSIS/ STAKEHOLDER ANALYSIS):

10. OBJECTIVES ACHIEVED SO FAR:

Completed:

**Data acquisition**

**Pre-processing**

Ongoing:

**Development of Deep Learning model**

**Result analysis**

11. WORK REMAINING TO BE DONE UNDER THE PROJECT:

**Fine tuning of developed Deep Learning model**

**Result analysis**

12. AGENCIES/ INSTITUTIONS/ DEPT.’S LIKELY TO BE INTERESTED IN THE PROBLEM, METHODOLOGY, RESULTS, ETC.

13. CONSTRAINTS, IF ANY:

14. INTERVENTION OF DST REQUESTED FOR:

15. DATE WHEN THIS REPORT WAS DISCUSSED WITH PROJECT TEAM/TARGET GROUP – AND BRIEF OUTCOME

DATE: SIGNATURE OF PI & CO-I