Integrating Regression Models and Climatological Data for Improved Precipitation Forecast in Southern India

J. Subha[1], S. Saudia[2]
Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli–12, Tamil Nadu, India[1, 2]

***Abstract*—Modern technologies like Artificial Intelligence (AI) and Machine Learning (ML) replicate intelligent human behavior and offer solutions in all domains, especially for human protection and disaster management. Nowadays, in both rural and urban areas, flood control is a serious issue to overcome the vast disaster to life and property. The work proposes to identify an appropriate ML based precipitation forecast model for the flood-prone southern states of India namely Tamil Nadu, Karnataka, and Kerala which receive most precipitation using the climatological information obtained from the NASA POWER platform. The work investigates the effectiveness of ML forecasting models: Multiple Linear Regression (MLR), Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR) and Ensemble (E) learning approaches of E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR in forecasting precipitation. The E-MLR-RFR model produces improved and most precise precipitation forecast in terms of Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and R^2 values. A higher precipitation forecast can be used to provide early warning about the possible flood in any region.**

Keywords*—*Ensemble models; machine learning models; rainfall precipitation forecast; R-squared value

I. Introduction

Flood control is an important matter of concern to confront casualties and rehabilitation [1]-[2]. Even though many advanced techniques have been utilized to control flood, concerned authorities still struggle to safeguard citizens and to quickly update estimates of the damage caused by floods. The incidents which happened in Chennai, Kerala, Assam and Bangalore due to continuous heavy rainfall during the monsoon season [3]-[7] show that protecting all lives and physical resources is a critical matter. In India, monsoon rainfall accounts for 80% of annual precipitation [8]-[9]. Floods which occurred in Kerala in the years 2018, 2019 and 2020 led to death of around 483, 121 and 104 people respectively and immersion of many villages [10]-[11]. The majority of people were then stranded in their houses without access to enough food and water.

Also, such flood hazards have affected most other countries like Bangladesh, Nepal, Pakistan, Afghanistan, Saudi Arabia and Iran [43]-[44]. Among the notable causes of flood, sudden and enduring heavy rain is the most important cause in all countries [12]. Forecast models published in [13]-[16], [19] are helpful to safeguard citizens, manage floods, operating reservoirs at critical situations by predicting floods and disseminating flood alerts. Hitherto, most flood forecasting systems relied either on monitored data or those retrieved from satellites [17]-[18], [20]-[21]. In [21]-[22], precipitation forecast is performed by interpreting such dynamically changing atmospheric data such as temperature, rainfall, humidity and wind direction. Predicting rainfall/ precipitation from this large volume of unstructured weather data [22] is not a facile task. So, the work proposes to identify suitable climatological data pre-processing stages and ML models for an automated warning of flood to the public and authorities based on a precipitation forecast. ML models are producing remarkable solutions in the domain of weather forecast and disaster management [45]-[48].

So, the paper proposes a study to identify stand-alone and ensemble Machine Learning models, MLR, SVR, DTR, RFR, E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR towards flood prediction via heavy precipitation forecast for the flood-prone south Indian states like Tamil Nadu, Karnataka and Kerala. These states which receive most heavy rainfall have been prone to major floods [8]-[9], [11]. This proposed work focuses on identifying a precipitation forecasting system for the states by integrating climatological data in the ML framework. The climatological data of the highest precipitation receiving geo-spatial locations of those states are captured from the NASA POWER platform [30]. The work also analyses the results of ML based flood/ precipitation forecasting models in literature prior to the selection of stand-alone ML models MLR, SVR, DTR, RFR and ensemble ML models, E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR to work with the climatological data for precipitation forecast. It also attempts the effectiveness of pairwise correlation for feature selection before selecting the independent features for training the ML models. The ensemble model, E-MLR-RFR produces the improved and highest R^2- value to precipitation forecast.

This paper is organized in five sections. Section II outlines the literature in terms of related climatological datasets, models and results predicted. Section III elaborates the data pre-processing, feature selection stages and the ML algorithms used for precipitation forecast. Section IV presents the results and discussion upon results and Section V concludes the paper.

II. Related Works

ML based papers, studies published by researchers to increase the accuracy of precipitation/ flood forecast are discussed in this section. Barrera-Animas, A.Y., et al. [23] presented a comparison of rainfall forecast models based on ML and Deep Learning (DL) architectures, Long-Short Term Memory (LSTM), Stacked-LSTM, Bidirectional-LSTM Networks, Extreme Gradient Boost (XGBoost), an ensemble of Gradient Boosting Regressor, Linear Support Vector Regression, and an Extra-trees Regressor to project hourly rainfall volumes using time-series data from five major cities of the United Kingdom (UK). The performance of the models is assessed using the assessment metrics, Root Mean Squared Error, Mean Absolute Error, and Root Mean Squared Logarithmic Error to identify the bidirectional-LSTM Network as a best rainfall forecast model out of all the models examined. Aftab, S., et al [24] used data mining approaches by identifying hidden patterns within the available elements of historical weather data to estimate the amount of rainfall. Velasco, L. C., et al. [25] forecasted the rainfall of Iligan city in Southern Philippines using Support Vector Regression Machine (SVRM) based on a 4-year and 17-month rainfall dataset collected using an Automated Rain Gauge (ARG). The forecasting model demonstrated a Mean Square Error (MSE) of 3.46. Khan, T. A., et al. (2019) [26] made classifier and regression models for the investigation of flash floods based on data gathered from the Kund Malir seashore by sensors. For forecasting, the dataset was subjected to Logistic Regression, Quadratic Support Vector Machine, K-Nearest Neighbors (K-NN), Exponential Gaussian Process Regression (GPR) and Ensemble Bagged Tree. The GPR outperformed all other methods with a minimal RMSE of 0.0002 and a prediction speed of 35000 observations per second. Abdullah, A. S., et.al. (2021) [28] used Seasonal Autoregressive Integrated Moving Average (SARIMA) and SVM for rainfall/ precipitation forecast in Bogor City, Indonesia. The SVM delivered an accurate result with a minimal Mean Absolute Percentage Error (MAPE) for predicting rainfall. Sreehari, E., et. al. (2018) [29] used MLR to forecast the amount of rainfall using the dataset from the Nellore district of Andhra Pradesh, India. The MLR approach produced more accurate findings for the amount of rainfall than Simple Linear Regression (SLR). De Castro, J.T., et al. (2013) [31] developed a technique for a flash flood warning system using Short Message Service (SMS) with improved warning information based on rising water level and water velocity. The regression equation was created based on velocity and water level data that was collected over a seven-day period since they were thought to be flash-flood causes. In order to provide registered users with an early warning, the system computes the present and future risk of flooding based on the model. Rezaeianzadeh, M., et al. (2014) [32] did flood flow forecasting at the outlet of the Khosrow Shirin watershed in the Fars Province of Iran using Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), MLR, and Multiple Nonlinear Regression (MNLR). The MNLR models outperformed the ANN, ANFIS, and MLR models with smaller RMSE values. Saha, A., et al. (2021) [33] assessed the performance of flood susceptibility (FS) mapping predictions in the Koiya River basin in Eastern India. Eight flood conditioning variables based on the topography and hydro-climatological conditions were used to create a flood inventory map using the novel ensemble approach of Hyperpipes (HP) [47] and Support Vector Regression (SVR). The ensemble technique produced a higher accuracy of 0.915. The summary of the related works in terms of ML models, precipitation and watershed datasets used and prediction results in terms of MAE/ MSE/ RMSE/ Accuracy is presented in Table I.

TABLE I. SUMMARY OF RELATED PUBLICATIONS

Authors / Years	Title and Study Regions	Models	Metrics
[23]	Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications – UK cities	LSTM, SLSTM, BLSTM, XGBoost, LSVR	RMSE-0.0084
[24]	Rainfall prediction using data mining techniques: A systematic literature review	ML, DL	-
[25]	Rainfall Forecasting using Support Vector Regression Machines – Southern, Philippines	SVRM	MSE-3.46
[26]	A comparison review based on classifiers and regression models for the investigation of flash floods – Kund Malir, Pakistan	LR, QSVM, KNN, EGPR	RMSE-0.0002
[28]	Comparison of SARIMA and SVM model for rainfall forecasting in Bogor city, Indonesia	SARIMA, SVM	RMSE - 63.25
[29]	Prediction of climate variable using multiple linear regression- AP, India	MLR, SLR	-
[31]	Flash flood prediction model based on multiple regression analysis for decision support system – Garang River, Semarang	Multiple Regression Technique	-
[32]	Flood flow forecasting using ANN, ANFIS and regression models- Shirin watershed, Iran	ANN, ANFIS, MLR, MNLR	R^2 – 0.81
[33]	Flood susceptibility assessment using novel ensemble of hyperpipes and support vector regression algorithms – Koiya River Basin - India	HP, SVR, HP-SVR, Ensemble Technique	Accuracy-0.915

From Table I, it is clear that knowing and evaluating variations in rainfall is essential for forecasting flood calamities. So, this paper proposes to use climatological data gathered from the NASAPOWER dataset [30] for the years from 2001 to 2020 in identifying a better ML based model to forecast precipitation in various flood-prone geographic areas across the southern states of the country namely Tamil Nadu, Karnataka and Kerala. The work also investigates the improvement in precipitation forecast via pairwise correlation in the feature selection stage of the ML workflow. An accurate forecast precipitation can be used to alarm associated flood and plan relevant precautionary measures.

III. Proposed System

The focus of this paper includes (i) gathering climatological data from the platform https://power.larc.nasa.gov/ [30] (ii) utilization of ML models in predicting the weather data-based day-wise and month-wise precipitation and (iii) evaluation of results. For the purpose of making an early flood warning from precipitation forecast in a southern region of India, the daily and monthly based climatological features are used in this work. Relevant climatological features are scaled using Standardization to the range from 0 to 1. Following this process, relevant independent features in the dataset are identified using pairwise correlation [52] and then subjected to ML regression models and ensemble models to forecast precipitation. Methodology of the work as illustrated in Fig. 1 is explained in upcoming sub-sections.

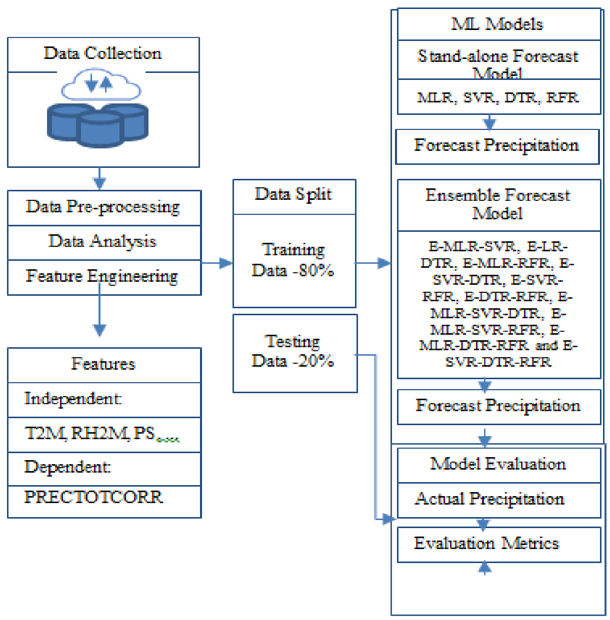


Fig. 1. Methodology of the proposed work.

A. Data Collection

The climatological data is acquired from the website, https://power.larc.nasa.gov/ [30] to forecast precipitation using the ML functions in scikit learn package of Python. This site provides solar and meteorological data sets from satellite observations for renewable energy and agricultural needs. For user-selected grids, the solar and meteorological characteristics are offered in monthly and daily basis. This climatological dataset is gathered for the three distinct southern Indian states, Tamil Nadu, Karnataka and Kerala of India. The day-wise and month-wise precipitation data is collected for the years from 2001 to 2020 in this region in Comma Separated Values [CSV] file format. The dataset for each region consists of 7305 records and 21 features. The details of the features in the dataset are mentioned in Table II.

TABLE II. Attributes in Datasets

Features Name	Description	Units	Range
YEAR	Year	Integer	2001-2020
DOY	Month/ Day of Year	Integer	1-366
PS	Surface pressure	(kPa)	15.81-31.98
TS	Earth skin temperature	C	0.92-22.95
T2M	Temperature at 2 meters	C	14.68-34.80
QV2M	Specific humidity at 2 meters	(g/kg)	1.82-22.40
RH2M	Relative humidity at 2 meters	%	21.54-41.12
WD2M	Wind direction at 2 meters	(degrees) ^{0}	8.83-25.23
WS2M	Wind speed at 2 meters	(m/s)	298.38-414.70
WD10M	Wind direction at 10 meters	(degrees) ^{0}	4.15-18.92
WS10M	Wind speed at 10 meters	(m/s)	21.75-95.88
T2MDEW	Dew/Frost point at 2 meters	C	92.53-94.27
T2M_MAX	Temperature at 2 meters maximum	C	0.48-7.52
T2M_MIN	Temperature at 2 meters minimum	C	0.78-9.41
WS2M_MAX	Wind speed at 2 meters maximum	(m/s)	0.60-3.90
WS2M_MIN	Wind speed at 2 meters minimum	(m/s)	20.94-336.88
T2M_RANGE	Temperature at 2 meters range	C	0.74-10.76
WS10M_MAX	Wind speed at 10 meters maximum	m/s	1.2-13.03
WS10M_MIN	Wind speed at 10 meters minimum	m/s	0.02-9.22
ALLSKY_ SFC_ LW_DWN	All sky surface longwave downward Irradiance	(w/m^2)	21.62-336.19
PRECTOTCORR	Precipitation corrected	(mm/day)	0-88.46

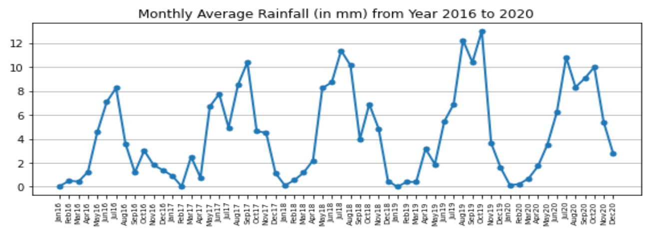
B. Data Pre-Processing

After data acquisition, data pre-processing is carried out to remove any null, empty or outlier data and to restore them with appropriate data for Machine Learning. Outliers are extreme data values which are out of observation ranges. The outliers/ missing values are rectified/ filled out to remove data irregularities and finally the entire data is transformed to the required format.

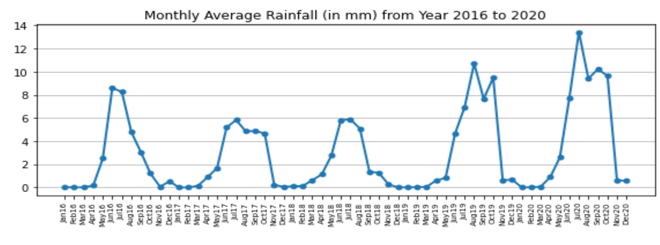
The climatological data obtained for the southern states, Tamil Nadu, Karnataka, and Kerala from [30] are devoid of missing values and outliers and so no pre-processing steps were necessary. The dataset is subjected to Data analysis and Feature Engineering to identify more relevant climatological features for training the different ML algorithms for forecasting rainfall. The later data analysis and feature engineering stages are detailed in the upcoming sub-sections.

C. Data Analysis

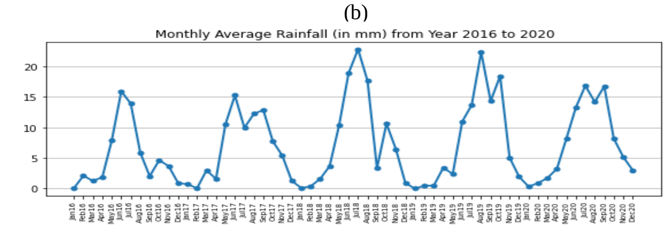
The climatological dataset used in the work is a high-resolution climatological data for a period of 20 years from 2001 to 2020. It is analyzed prior feature engineering using visualization tools of Python packages namely seaborn and matplotlib. The mean precipitation of Tamil Nadu, Karnataka, and Kerala over a period of time from 2016 to 2020 is shown in Fig. 2(a) to 2(c).



(a)



(b)



(c)

Fig. 2. Line plots showing average monthly precipitation from 2016 to 2020 (a) Tamil Nadu (b) Karnataka (c) Kerala.

The Fig. 2(a) clearly shows nine peak points from the climatological data of Tamil Nadu for the months July 2016, June 2017, September 2017, July 2018, October 2018, August 2019, October 2019, July 2020 and October 2020. Fig. 2(b) shows seven rainfall peak points of Karnataka which occurred in the months June 2016, July 2017 and 2018, August 2019, October 2019, July 2020 and September 2020.

Fig. 2(c) clearly shows ten peak precipitation points of Kerala in June and October of 2016, June and September of 2017, July and October of 2018, August and October of 2019 and July and September of 2020. Fig. 2(a) to (c) clearly indicate that there is maximum probability of precipitation and hence flood in monsoon season. Floods in these states have occurred during these high precipitation months [10], [66]-[67]. So, these months need special flood attention.

D. Feature Engineering

The goal of feature engineering in ML is to identify relevant independent features to make the models perform better. Scaling is one of the main goals of feature engineering which identifies the most pertinent quartiles of the data. In this work, the independent features in the respective ranges as shown in Table II are scaled using Standardization [49] to the range from 0 to 1. The scaling operation used to find the new independent feature value, x_{new} is shown mathematically in Eq. (1).

$$x_{new} = \frac{x-\mu}{\sigma} \tag{1}$$

where x is the independent feature, $\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$ is the mean of the N values of the independent feature x_i and σ is the standard deviation represented mathematically as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$.

Correlation analysis helps to collect essential features from the dataset. First, the correlation coefficients between independent features are determined and recorded in a table called a correlation matrix. The correlation between two features is represented in each cell of the correlation table from -1 to 1. If the value is positive, then there is a normal correlation; while bigger positive values indicate a stronger correlation between features.

There is an inverse correlation when the matrix values are negative. Relationship between feature pairs in a dataset is determined using the corr() function in python as a heat map in Fig. 3.The correlation heat map for Tamil Nadu state is shown in Fig. 3. It clearly demonstrates correlations between independent features with same cell colour and value. The feature T2M is highly correlated with TS, T2M_MAX. Also, the feature T2M_DEW is highly correlated with TS, T2M_RANGE, T2M_MIN, QV2M, RH2M ALLSKY_SFC_LW_DWN etc.

The range of correlation from 1 to -1 is shown by the colour intensity variations from green to blue. An independent feature is chosen from each pairwise correlated group to be included in model design phase [52]. This helps to reduce the number of features in the dataset which in turn can improve the performance of ML modeling [49] - [53]. The effectiveness and interpretability of model can be increased by the new collection of uncorrelated features.

Following the feature selection stage, the dataset for Tamil Nadu state has the features: RH2M, PS, WS10M_MAX, ALLSKY_SFC_LW_DWN and WD10M. The feature engineered dataset with the relevant set of independent and dependent features as shown in Table III is subjected to the model building stage.

TABLE III. FEATURES OBTAINED FROM THE FEATURE ENGINEERING STAGE

State	Tamil Nadu	Karnataka	Kerala
Independent Features	RH2M, PS, WS10M_MIN, ALLSKY_SFC_LW_DWN, WD10M	RH2M, PS, WD2M, WS10M_MIN, WD10M	T2M_MIN, QV2M, RH2M, PS, WS10M_MAX, WS10M_MIN, WD10M
Dependent Feature	PRECTOTCORR	PRECTOTCORR	PRECTOTCORR

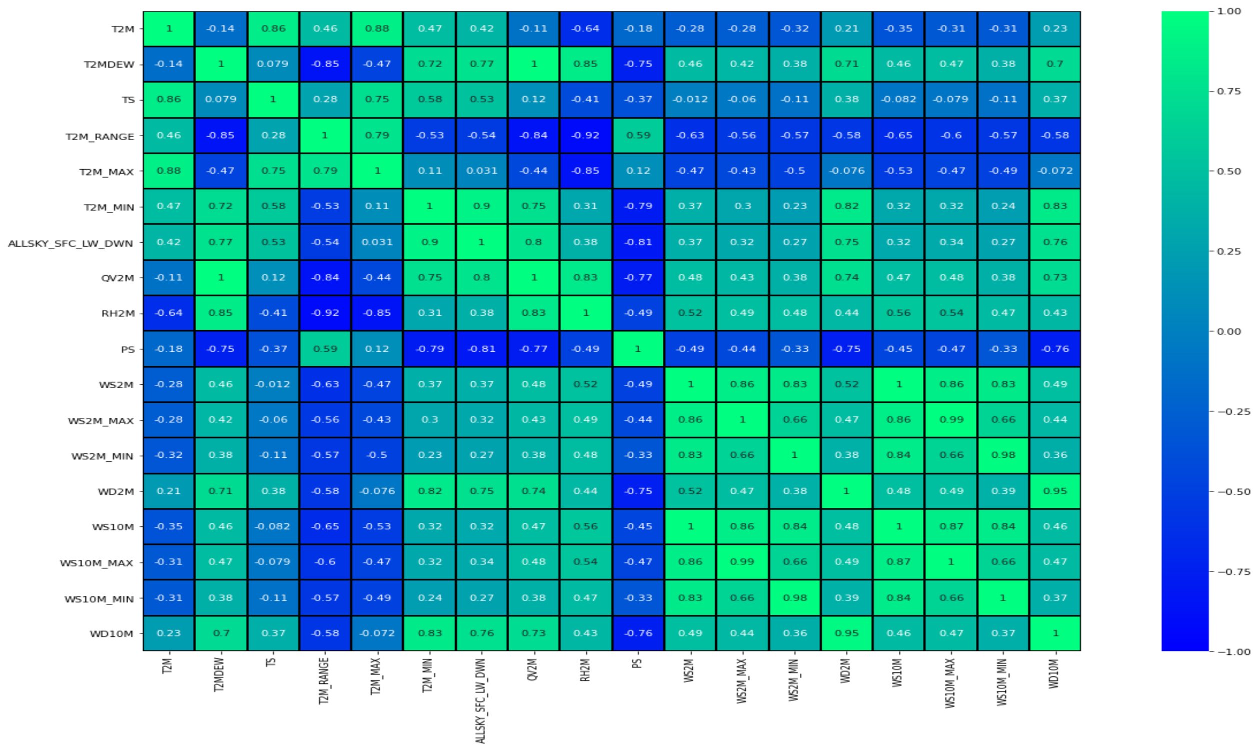


Fig. 3. Correlation heat map for Tamil Nadu state.

E. *Model Building*

This work proposes to identify the effectiveness of the standalone ML models: MLR, SVR, DTR, RFR and ensemble ML models: E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR in forecasting precipitation using the independent features obtained from the NASAPOWER dataset for the flood-prone south Indian states of Tamil Nadu, Karnataka and Kerala. The dataset with features as in Table III is split into training and testing datasets in the ratio, 8:2 and the training dataset is subjected to the training phases of the above mentioned standalone and ensemble ML models [36]-[37]. In ensemble models, the final prediction is made by averaging the results of all the base models. General training phase of these ML and ensemble algorithms are briefed as follows:

1) Machine learning models: Machine Learning (ML) [38]-[41] is the science of creating regression/ classification models using algorithms that can draw knowledge from prior instances. The general data flow diagram of a ML based regression algorithm for forecasting rainfall is as shown in Fig. 4. Fig. 4 depicts the typical data flow diagram of the regression algorithm where the independent features from the dataset in Table III are subjected to the training phase of the ML algorithms, both stand-alone algorithms: MLR, SVR, DTR, RFR and ensemble algorithms: E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR to produce the model to forecast precipitation. The precipitation forecast is obtained from these models using the test data. The working of different ML and ensemble algorithms are briefed in sub-sections below.

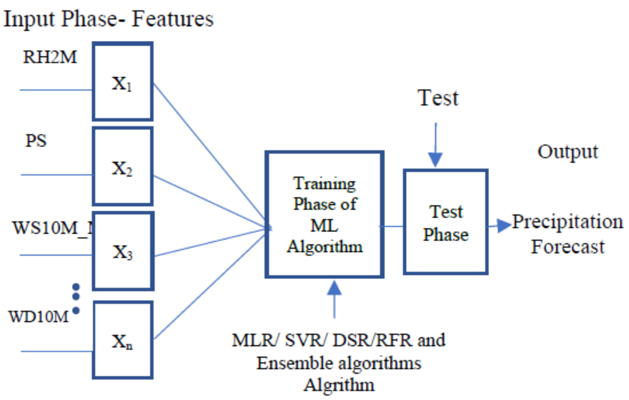


Fig. 4. General data flow diagram of regression algorithm.

2) Multi linear regression: Multiple Linear Regression [6], [41] determines the relation between dependent feature and many independent features to forecast the values of the continuous dependent feature. In MLR, the relationship between the independent features, x and dependent feature, y is modeled in the training phase as a linear equation shown in Eq. (2) by minimizing the sum of squares of error between the actual and predicted dependent feature values or residuals using Least Squares optimization [54]-[55].

$$y_{pi} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \qquad (2)$$

Here the indices $1,2,3, \dots, n$ correspond to the 'n' independent features, y_{pi} is the dependent feature, b_0 is the y-intercept, and b_1 , b_2 ,...., b_n are the coefficients of the independent features x_1 , x_2 ,…, x_n respectively. The illustration is shown in Fig. 5. In this work, MLR is trained using the training set Tamil Nadu, Karnataka and Kerala states and the model is used to forecast daily and monthly precipitation.

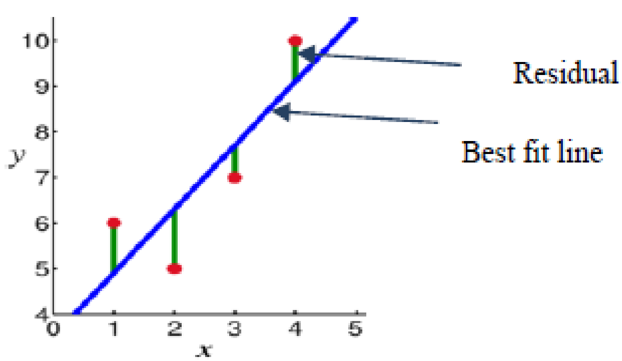


Fig. 5. General diagram for MLR [56].

3) Support vector regression: Support Vector Regression identifies a hyperplane in an n-dimensional space of the independent features, x in the training dataset as a model to predict the values of the dependent feature, y_{pi} . The hyperplane model obtained after training as shown in Eq. (3) has the largest margin between the support vectors, ξ *and* ξ *around a marginalized $'\varepsilon'$ space. The $'\varepsilon'$ space is +ε and -ε from the hyperplane.

$$y_{pi} = f(x) = bx + c = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*). K(x_i, x) + c \quad (3)$$

Here b is the weight vector corresponding to the independent feature x in terms of the Lagrange multipliers $\alpha_i, \alpha_i^*$and c is the bias term which are obtained in the training phase by minimizing the objective function, $\frac{1}{2}||b||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i{}^*)$ using quadratic optimization [57]-[59]. K is the Kernal function which transforms x to higher dimension and C is the penalty parameter of the model and N is the size of the training dataset. The illustration of the parameters are shown in Fig. 6. In this work, the SVR model is used for predicting continuous values of precipitation for Tamil Nadu, Karnataka and Kerala on a daily and monthly basis.

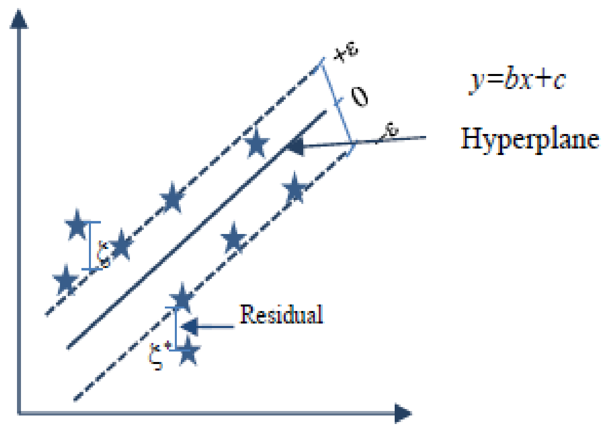


Fig. 6. General diagram for SVR [59].

4) Decision tree regression: Regression issues can be solved using the Decision Tree Regressor (DTR) as shown in Fig. 7 [60], [63]. Starting with the root node with all the records in the training dataset, a decision tree is built. The tree divides into left and right child nodes based on a condition check on an independent feature values with least Mean Square Error to contain subsets of the training dataset. Mean Square Error is the difference between the predicted values of the dependent feature and its original target value.

The child nodes are further subdivided into their children nodes and thus become the parent nodes of next level. Each branch denotes the outcome of a test and each leaf node denotes the final outcome. The set of all conditions until different leaf nodes corresponds to the regression model. The predicted output is obtained as the average of the dependent feature values of all records in the leaf node. DTR model trained using the training set from Tamil Nadu state to forecast daily precipitation is shown in Fig. 7.

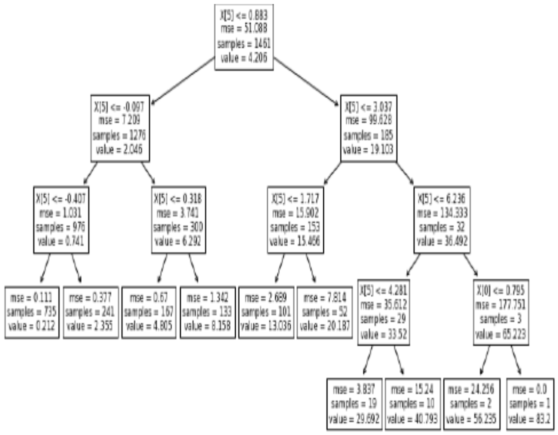


Fig. 7. Sample DTR with five levels drawn from climatological day-wise precipitation data of Tamil Nadu state.

5) Random forest regression: Random Forest Regressor (RFR) as shown in Fig. 8 has a large number of decision tree regressors trained from the subsets of the training dataset [65][27]. It is a bagging ensemble approach which employs aggregated decision trees that operate concurrently without interacting with one another and produces the regression output as the average of outputs from all decision tree regressors [61], [64]. One out of ten DTRs used in the RFR model trained using the training set from Tamil Nadu state to forecast daily precipitation is shown in Fig. 9.

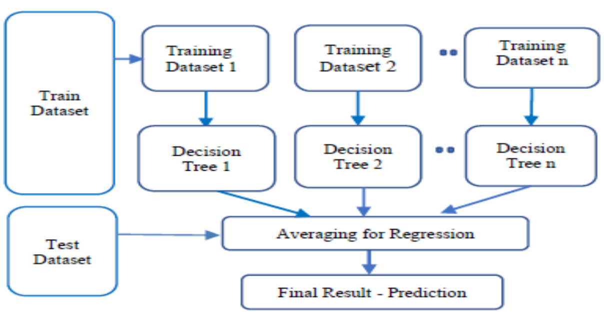


Fig. 8. The random forest regression ensemble.

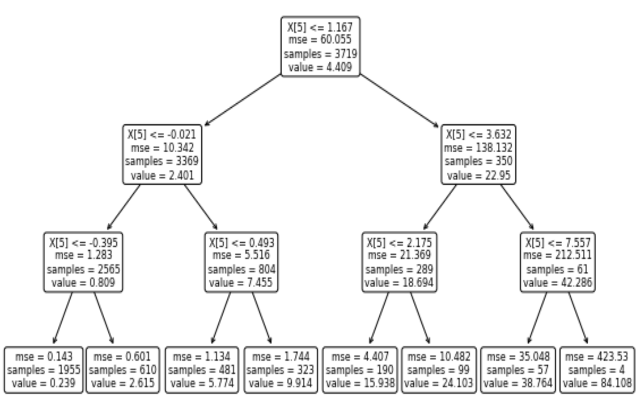


Fig. 9. One sample DTRs of RFR drawn from climatological data for Tamil Nadu state.

F. *Ensemble Models*

In Machine Learning (ML), the use of multiple models or algorithms to increase prediction reliability is referred to as ensemble learning [61]. The fundamental idea behind ensemble approaches is to achieve improved results using Eq. (4) by integrating results from many models than from a single model [61]-[62]. The average of results from the numerous regression models is the final prediction. In this work, the ensemble techniques, E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR are used to forecast precipitation.

$$Final\ Prediction = \frac{\sum_{j=1}^{m}(Prediction\ from\ Model\ j)}{m} \tag{4}$$

where $j = 1$ to m and m is the number of models.

The ensemble approaches used in this work are bagging/ averaging approaches where arbitrary subsets of the training data are trained in several base algorithms: MLR, SVR, DTR, RFR and the predictions from each model are combined to get a final prediction. In the ensemble approach, E-MLR-SVR, the outputs of the basic regressors, MLR and SVR are combined to get the final output. Integrating stand-alone models yields better outcomes than using a stand-alone model. The climatological datasets from Tamil Nadu, Karnataka and Kerala are trained using the stand-alone ML algorithms and ensemble learning approaches to predict daily and monthly rainfall in the states of Tamil Nadu, Karnataka and Kerala. The performance of these models are assessed in terms of MAE, MSE, RMSE and R^2 values. Section IV analyses the results of these models in terms of regression metrics.

IV. Experimental Results and Analysis

The performance evaluation metrics, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and R^2 (R- squared) values are used to assess precipitation forecasting models [34]-[35]. MAE refers to the average of the absolute error difference between predicted value, y_{pi} and actual value, y_i as defined in Eq. (5). RMSE defined in Eq. (6) is the square root of the mean square error (MSE). The percentage of the dependent feature's fluctuation that can be predicted from the independent feature is known as R^2 value. It is defined in Eq. (7).

$$MAE = \frac{\sum_{i=1}^{N}|y_i - yp_i|}{N} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_{pi})^2} \tag{6}$$

where $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_{pi})^2$, N is the total number of observations/rows in the test data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - y_{pi})^2}{\sum_{i=1}^{N}(y_i - \bar{y}_{pi})^2} \tag{7}$$

where y_i is actual value of i^{th} observation, y_{pi} is predicted value of i^{th} observation, $\bar{y}_{pi}$ is the average of predicted values and N is the number of observations/rows.

A. Results and Analysis

The testing experiments are conducted on the test dataset obtained from the 20% climatological data of Tamil Nadu, Kerala and Karnataka states in an Intel Core™ i7-7500 CPU with 2.70 GHZ speed and 16GB RAM using the numerical and ML packages of Python. The results obtained in terms of MAE, MSE, RMSE and R2 values from the stand-alone ML algorithms: MLR, SVR, DTR, RFR and ensemble algorithms: E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR are tabulated in the Tables IV and V for day-wise and month-wise predictions. It is found from Tables IV and V that E-MLR-RFR produces improved precipitation forecast than other models in terms of MAE, MSE, RMSE and R2 values for Tamil Nadu, Karnataka and Kerala states and also than the models in [25], [28] and [32] in Table I. The minimal RMSE and maximal R2 value for day-wise precipitation forecast from the E-MLR-RFR model are 0.11, 0.2, 0.1 and 0.9997, 0.999, 0.999 for Tamil Nadu, Karnataka and Kerala states respectively. Also the minimal RMSE and maximal R^2 value for month-wise precipitation forecast from the E-MLR-RFR model are 0.06, 0.08, 0.31 and 0.9996, 0.999, 0.999 for Tamil Nadu, Karnataka and Kerala states respectively. The performance comparisons of the precipitation predictions made by different models: stand-alone algorithms: MLR, SVR, DTR, RFR and ensemble algorithms: E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR are respectively shown in the line diagrams from Fig. 10 to Fig. 12 for Karnataka state. Minimal error is noted from all the line plots in terms of predicted and actual monthly precipitation.

TABLE IV. MAE, MSE, RMSE AND R2 VALUES OF VARIOUS REGRESSION METHODS FOR TAMIL NADU (TN), KARNATAKA (KA) AND KERALA (KL) – DAY-WISE PREDICTIONS

	MAE			MSE			RMSE			R^2 value		
Models	TN	KA	KL	TN	KA	KL	TN	KA	KL	TN	KA	KL
MLR	0.07	0.54	0.25	0.01	0.53	0.16	0.10	0.73	0.41	0.999	0.994	0.99
SVR	0.31	0.62	0.88	6.90	3.35	42.78	2.62	1.83	6.54	0.864	0.966	0.75
DTR	1.43	1.78	2.60	6.08	8.32	24.42	2.46	2.88	4.94	0.88	0.91	0.85
RFR	0.49	0.79	0.96	0.73	1.98	5.24	0.85	1.41	2.28	0.985	0.98	0.96
E-MLR-SVR	0.20	0.14	0.16	0.35	0.25	0.47	0.59	0.50	0.68	0.99	0.997	0.99
E-MLR-DTR	0.73	0.81	1.28	1.31	1.70	4.94	1.14	1.30	2.22	0.97	0.98	0.96
E-MLR-RFR	**0.006**	**0.004**	**0.009**	**0.012**	**0.008**	**0.01**	**0.11**	**0.20**	**0.10**	**0.9997**	**0.999**	**0.999**
E-SVR-DTR	0.82	0.86	1.34	2.00	2.11	6.81	1.41	1.45	2.60	0.958	0.975	0.95
E-SVR-RFR	0.20	0.14	0.16	0.37	0.25	0.50	0.61	0.50	0.70	0.99	0.997	0.99
E-DTR-RFR	0.73	0.81	1.28	1.44	1.70	4.92	1.20	1.30	2.22	0.96	0.98	0.96
E-MLR-SVR-RFR	0.13	0.09	0.10	0.17	0.11	0.20	0.41	0.33	0.45	0.996	0.998	0.99
E-MLR-SVR-DTR	0.54	0.57	0.89	0.89	0.93	3.02	0.94	0.96	1.73	0.981	0.989	0.98
E-MLR-DTR-RFR	0.49	0.54	0.85	0.62	0.74	2.21	0.78	0.86	1.48	0.987	0.99	0.985
E-SVR-DTR-RFR	0.54	0.57	0.89	0.93	0.92	3.04	0.96	0.96	1.74	0.98	0.98	0.98

TABLE V. MAE, MSE, RMSE AND R2 VALUES OF VARIOUS REGRESSION METHODS FOR TAMIL NADU (TN), KARNATAKA (KA) AND KERALA (KL) – MONTH-WISE PREDICTIONS

	MAE			MSE			RMSE			R^2 value		
Models	TN	KA	KL	TN	KA	KL	TN	KA	KL	TN	KA	KL
MLR	0.60	1.31	1.25	0.45	1.72	2.05	0.67	1.31	1.43	0.96	0.95	0.95
SVR	0.76	1.19	1.69	1.06	3.41	6.40	1.03	1.84	2.54	0.92	0.91	0.86
DTR	0.91	1.36	1.66	1.58	3.23	5.26	1.25	1.79	2.29	0.88	0.92	0.89
RFR	0.63	0.87	1.24	0.63	1.07	2.54	0.79	1.04	1.59	0.95	0.97	0.94
E-MLR-SVR	0.95	1.48	2.24	1.67	5.97	9.30	1.29	2.44	3.06	0.87	0.85	0.81
E-MLR-DTR	0.30	0.53	0.75	0.16	0.55	1.23	0.40	0.74	1.10	0.988	0.98	0.97
E-MLR-RFR	**0.02**	**0.03**	**0.09**	**0.004**	**0.007**	**0.09**	**0.06**	**0.08**	**0.31**	**0.9996**	**0.999**	**0.998**

E-SVR-DTR	1.04	1.58	2.73	2.08	6.49	14.87	1.44	2.54	3.85	0.85	0.84	0.71
E-SVR-RFR	0.96	1.45	2.33	1.77	5.63	11.06	1.33	2.37	3.32	0.87	0.86	0.78
E-DTR-RFR	0.31	0.53	0.80	0.19	0.57	1.86	0.44	0.75	1.36	0.99	0.98	0.96
E-MLR-SVR-RFR	0.64	0.97	1.53	0.78	2.53	4.75	0.88	1.59	2.18	0.94	0.93	0.90
E-MLR-SVR-DTR	0.69	1.05	1.82	0.92	2.88	6.60	0.96	1.69	2.57	0.93	0.93	0.87
E-MLR-DTR-RFR	0.21	0.36	0.54	0.08	0.26	0.93	0.28	0.50	0.96	0.99	0.99	0.98
E-SVR-DTR-RFR	0.70	1.04	1.88	0.98	2.74	7.66	0.99	1.45	2.76	0.92	0.93	0.85

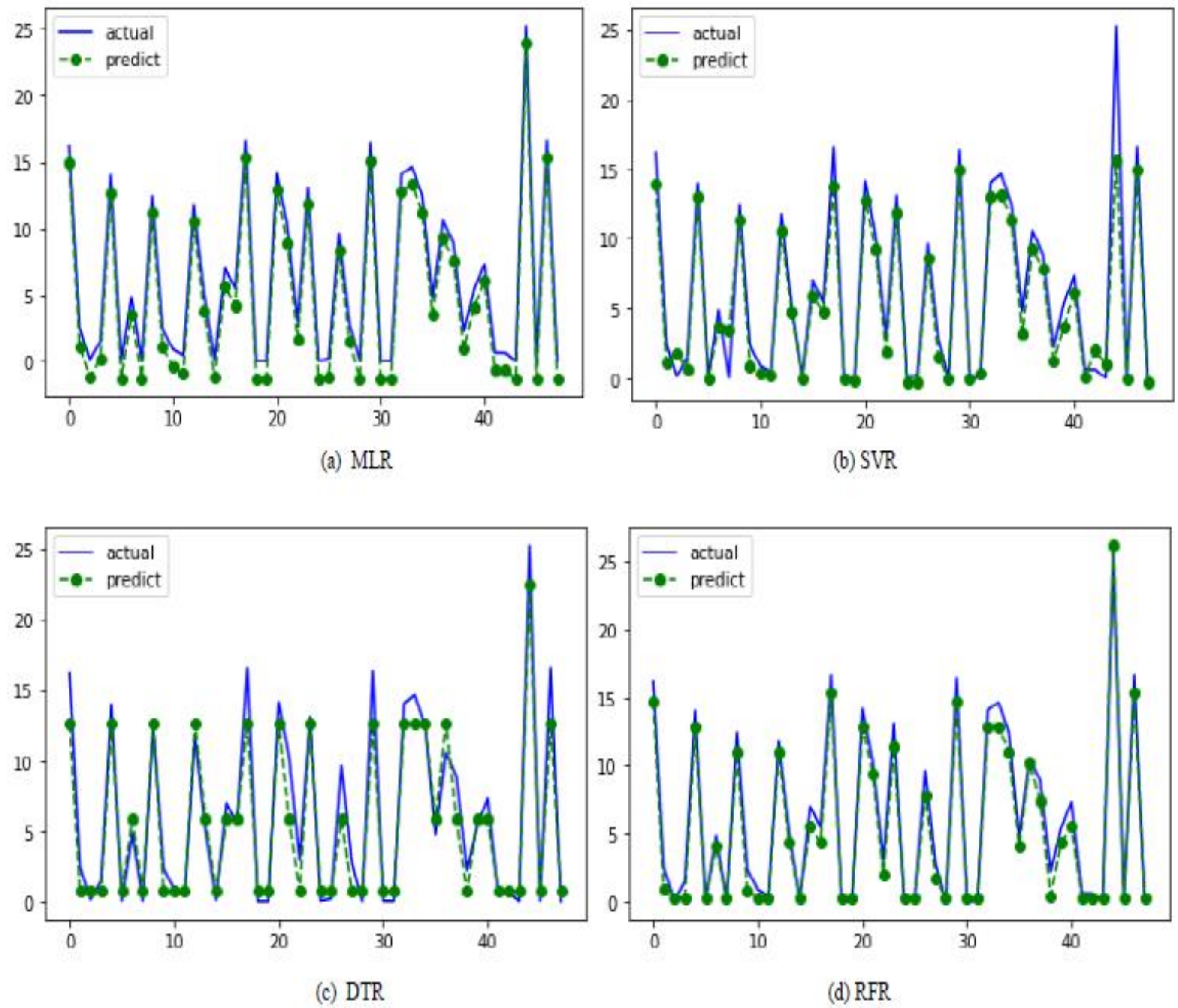
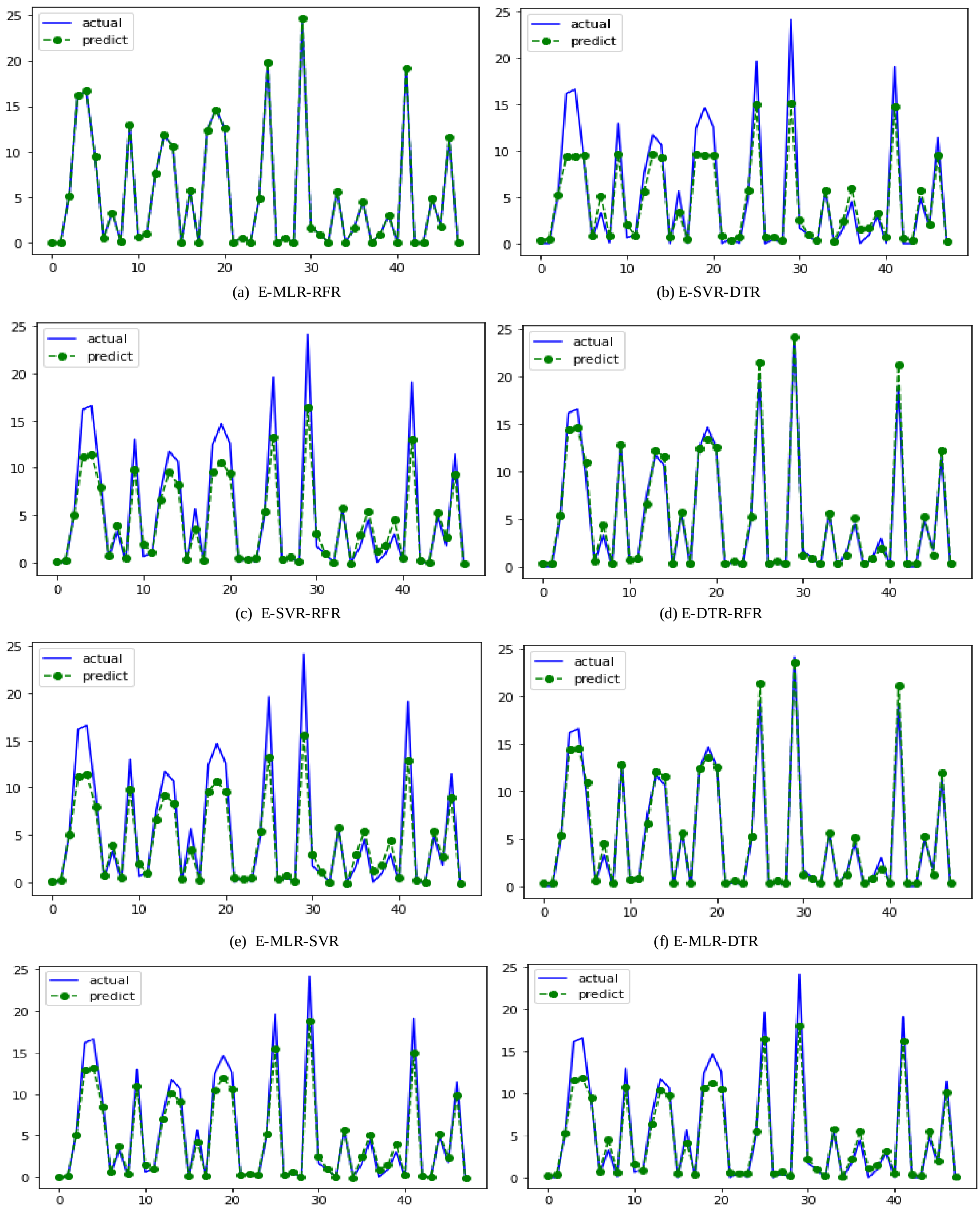


Fig. 10. Line plots showing actual and predicted monthly precipitation of Karnataka (a) MLR (b) SVR (c) DTR (d) RFR.



(a) E-MLR-RFR

(b) E-SVR-DTR

(c) E-SVR-RFR

(d) E-DTR-RFR

(e) E-MLR-SVR

(f) E-MLR-DTR

(g) E-MLR-SVR-RFR.

(h) E-MLR-SVR-DTR.

Fig. 11. Line plots showing actual and predicted monthly precipitation of Karnataka (a) E-MLR-RFR (b) E-SVR-DTR (c) E-SVR-RFR (d) E-DTR-RFR (e) E-MLR-SVR (f) E-MLR-DTR (g) E-MLR-SVR-RFR and (h) E-MLR-SVR-DTR.

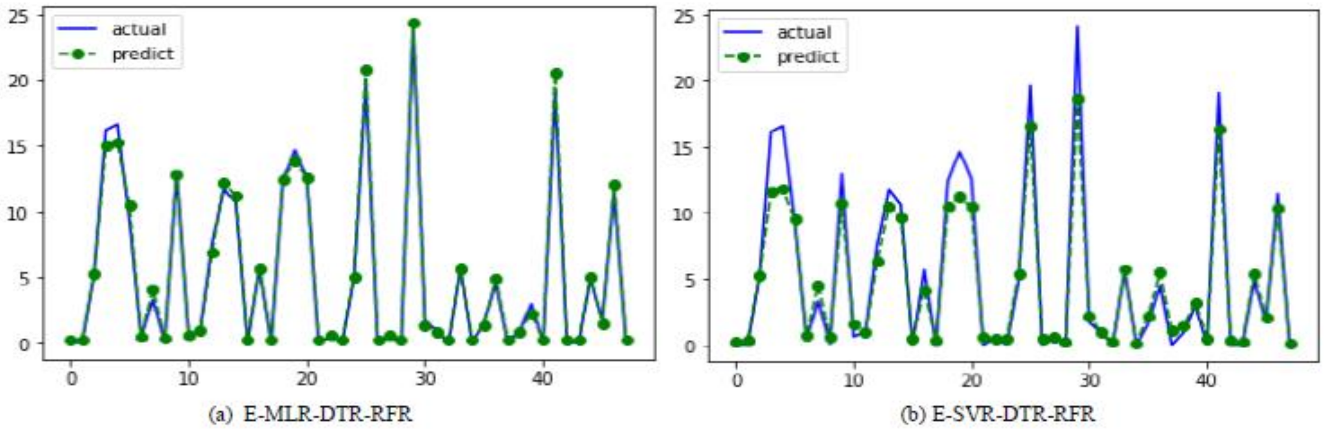


Fig. 12. Line plots showing actual and predicted monthly precipitation of Karnataka (a) E-MLR-DTR-RFR (b) E-SVR-DTR-RFR.

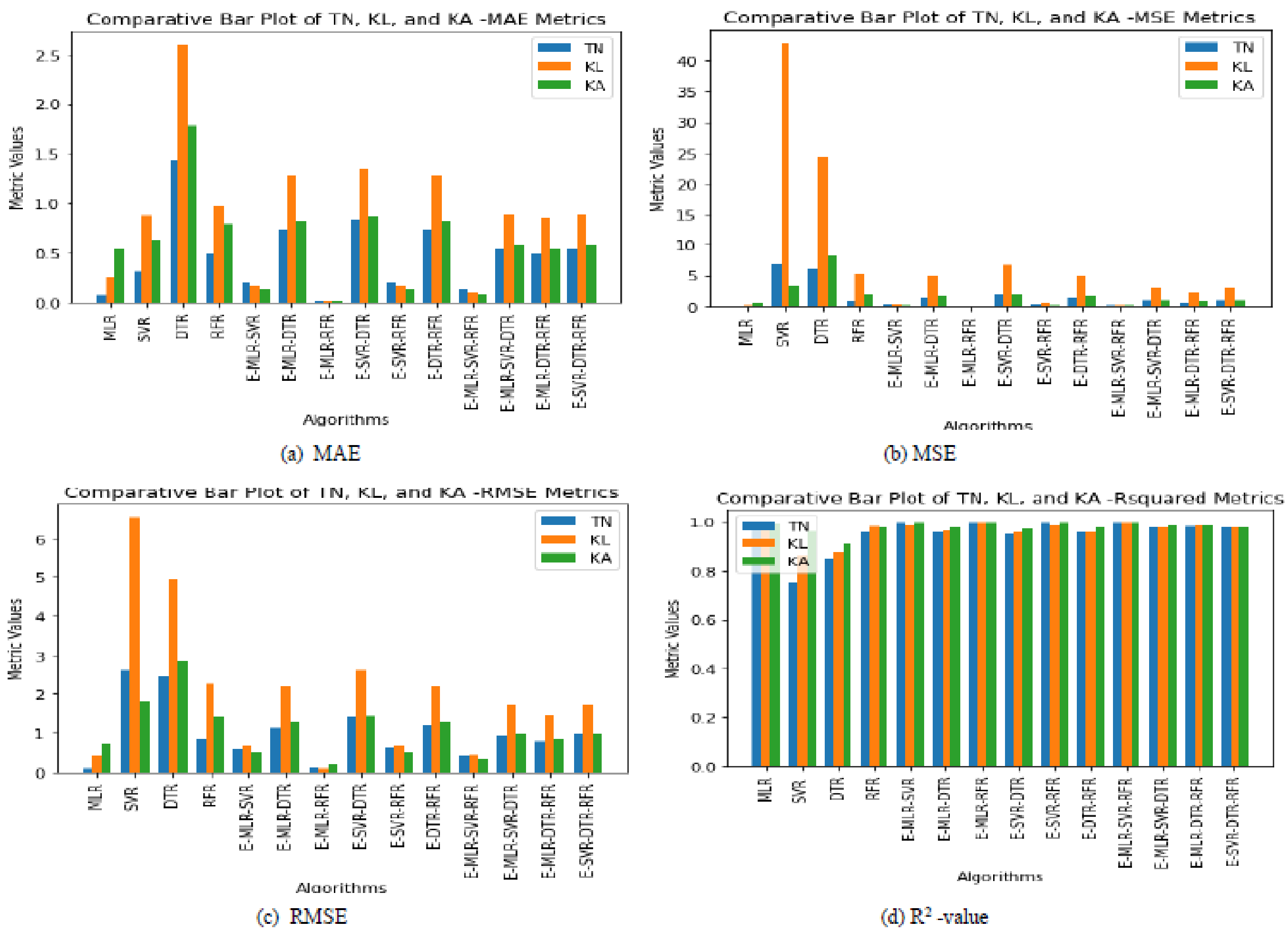


Fig. 13. Comparison of (a) MAE (b) MSE (c) RMSE (d) R^2-values for Tamil Nadu (TN), Karnataka (KA) and Kerala (KL).

A barplot comparison of the evaluation metrics, MAE, MSE, RMSE and R^2-values is also made between the states of Tamil Nadu, Karnataka and Kerala in Fig. 13. The ensemble model E-MLR-RFR fared better than other models. Also, the climatological data from Karnataka produces lesser error when compared to the data from Kerala and Tamil Nadu for all models in the comparative study.

Heavy rainfall is one of the main reasons for flooding. A buildup of water in low-lying places can result in rivers overflowing their banks when rainfall surpasses a specific threshold, often 35.6 mm or more as reported in [42] and shown in Table VI. So, the heavy rainfall predictions made from the best model, E-MLR-RFR proposed from the work to provide flood warnings in a specific location to reduce serious damage to both the environment and human societies.

TABLE VI. RAINFALL RANGE-FLOOD RANGE IS HIGHLIGHTED BOL

Description	Rainfall amount (mm/day)
Very Light Rain	0.1 -2.4
Light Rain	2.5 - 7.5
Moderate Rain	7.6- 35.5
Rather Heavy Rain	**35.6-64.4**
Heavy Rain	**64.5 - 124.4**
Very Heavy Rain	**124.5 - 244.4**
Extremely Heavy Rain	**>244.4**

V. CONCLUSION

The work has identified an appropriate ML based precipitation forecast model for the flood-prone southern states of India namely Tamil Nadu, Karnataka, Kerala which receive most precipitation using climatological data obtained from the NASA POWER platform. The precipitation forecast is proposed to alarm flood. The work investigated the effectiveness of ML forecasting models: Multiple Linear Regression (MLR), Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR) and ensemble approaches E-MLR-SVR, E-MLR-DTR, E-MLR-RFR, E-SVR-DTR, E-SVR-RFR, E-DTR-RFR, E-MLR-SVR-DTR, E-MLR-SVR-RFR, E-MLR-DTR-RFR and E-SVR-DTR-RFR in forecasting precipitation. The E-MLR-RFR model has produced improved and precise precipitation in terms of Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and R^2 values. The higher precipitation forecast from E-MLR-RFR can be used to provide early warning about the possible flood in any region.

ACKNOWLEDGMENT

The data is obtained from the website: https://power.larc.nasa. gov/ [30]. The website is associated with NASA Langley Research Center's (LaRC) POWER Project -a NASA Earth Science/Applied Science Program.

REFERENCES

[1] Lamond, J., Stanton-Geddes, Z., Bloch, R., & Proverbs, D. (2013). Cities and Flooding: Lessons in resilience from case studies of integrated urban flood risk management. CIB.

[2] Nkwunonwo, U. C., Whitworth, M., & Baily, B. (2016). A review and critical analysis of the efforts towards urban flood risk management in the Lagos region of Nigeria. *Natural hazards and earth system sciences*, *16*(2), 349-369.

[3] Svetlana, D., Radovan, D., & Ján, D. (2015). The economic impact of floods and their importance in different regions of the world with emphasis on Europe. *Procedia Economics and Finance*, *34*, 649-655.

[4] Chaluvadi, R., Varikoden, H., Mujumdar, M., Ingle, S. T., & Kuttippurath, J. (2021). Changes in large-scale circulation over the Indo-Pacific region and its association with 2018 Kerala extreme rainfall event. *Atmospheric Research*, *263*, 105809.

[5] [5] Bashir, O. O., Oludare, A. H., Johnson, O. O., & Aloysius, B. (2012). Floods of fury in Nigerian cities. *Journal of Sustainable Development*, *5*(7), 69.

[6] Nair, P. J., Varikoden, H., Francis, P. A., Chakraborty, A., & Pandey, P. C. (2021). Atmospheric moisture as a proxy for the ISMR variability and associated extreme weather events. *Environmental Research Letters*, *16*(1), 014045.

[7] Few, R. (2003). Flooding, vulnerability and coping strategies: local responses to a global threat. *Progress in Development studies*, *3*(1), 43-58.

[8] Subha, J., & Saudia, S. (2022). An Exploratory Data Analysis on SDMR Dataset to Identify Flood-Prone Months in the Regional Meteorological Subdivisions. In *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2022* (pp. 595-617). Singapore: Springer Nature Singapore.

[9] Kumar, V., Jain, S. K., & Singh, Y. (2010). Analysis of long-term rainfall trends in India. *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, *55*(4), 484-496.

[10] Wikipedia contributors. (2023). 2018 Kerala floods. *Wikipedia*. https://en.wikipedia.org/wiki/2018_Kerala_floods.

[11] Mishra, V., & Shah, H. L. (2018). Hydroclimatological perspective of the Kerala flood of 2018. *Journal of the Geological Society of India*, *92*(5), 645-650.

[12] Kuttippurath, J., Murasingh, S., Stott, P. A., Sarojini, B. B., Jha, M. K., Kumar, P., ... & Pandey, P. C. (2021). Observed rainfall changes in the past century (1901–2019) over the wettest place on Earth. *Environmental Research Letters*, *16*(2), 024018.

[13] Magar, R. B., & Jothiprakash, V. (2011). Intermittent reservoir daily-inflow prediction using lumped and distributed data multi-linear regression models. *Journal of earth system science*, *120*(6), 1067-1084.

[14] Philipp, A., Schmitz, G. H., Krausse, T., Schütze, N., & Cullmann, J. (2008). Flash flood forecasting combining meteorological ensemble forecasts and uncertainty of initial hydrological conditions. *Australasian Journal of Water Resources*, *12*(3), 257-267.

[15] Tate, E., & Cauwenberghs, K. (2005). An innovative flood forecasting system for the Demer basin: A case study. *International Journal of River Basin Management*, *3*(3), 163-167.

[16] Tsakiri, K., Marsellos, A., & Kapetanakis, S. (2018). Artificial neural network and multiple linear regression for flood prediction in Mohawk River, New York. *Water*, *10*(9), 1158.

[17] Khalaf, M., Alaskar, H., Hussain, A. J., Baker, T., Maamar, Z., Buyya, R., ... & Al-Jumeily, D. (2020). IoT-enabled flood severity prediction via ensemble machine learning models. *IEEE Access*, *8*, 70375-70386.

[18] Khan, T. A., Alam, M. M., Shahid, Z., & Su'Ud, M. M. (2020). Investigation of Flash Floods on early basis: A Factual Comprehensive review. *IEEE Access*, *8*, 19364-19380.

[19] Al-Dabbagh, A. W., Hu, W., Lai, S., Chen, T., & Shah, S. L. (2018). Toward the advancement of decision support tools for industrial facilities: Addressing operation metrics, visualization plots, and alarm floods. *IEEE Transactions on Automation Science and Engineering*, *15*(4), 1883-1896.

[20] Zaji, A. H., Bonakdari, H., & Gharabaghi, B. (2018). Applying upstream satellite signals and a 2-D error minimization algorithm to advance early warning and management of flood water levels and river discharge. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(2), 902-910.

[21] Manandhar, S., Dev, S., Lee, Y. H., Meng, Y. S., & Winkler, S. (2019). A data-driven approach for accurate rainfall prediction. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(11), 9323-9331.

[22] Chen, H., Chandrasekar, V., Cifelli, R., & Xie, P. (2019). A machine learning system for precipitation estimation using satellite and ground radar network observations. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(2), 982-994.

[23] Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applicatins*, *7*,100204.

[24] Aftab, S., Ahmad, M., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction using data mining techniques: A systematic literature review. *International journal of advanced computer science and applications*, *9*(5).

[25] Velasco, L. C., Aca-ac, J. M., Cajes, J. J., Lactuan, N. J., & Chit, S. C. (2022). Rainfall Forecasting using Support Vector Regression Machines. *International Journal of Advanced Computer Science and Applications*, *13*(3).

[26] Khan, T. A., Alam, M., Kadir, K., Shahid, Z., & Mazliham, M. S. (2019). A comparison review based on classifiers and regression models for the investigation of flash floods. *Editorial Preface from the Desk of Managing Editor*, *10*(9), *352-359*.

[27] Meenal, R., Michael, P. A., Pamela, D., & Rajasekaran, E. (2021). Weather prediction using random forest machine learning model. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(2), 1208-1215.

[28] Abdullah, A. S., Ruchjana, B. N., & Jaya, I. G. N. M. (2021). Comparison of SARIMA and SVM model for rainfall forecasting in Bogor city, Indonesia. In *Journal of Physics: Conference Series* (Vol. 1722, No. 1, p. 012061). IOP Publishing.

[29] Sreehari, E., & Srivastava, S. (2018). Prediction of climate variable using multiple linear regression. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)* (pp. 1-4). IEEE.

[30] *NASA POWER*, https://power.larc.nasa.gov/

[31] de Castro, J. T., Salistre Jr, G. M., Byun, Y. C., & Gerardo, B. D. (2013). Flash flood prediction model based on multiple regression analysis for decision support system. In *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 2, pp. 23-25).

[32] Rezaeianzadeh, M., Tabari, H., Arabi Yazdi, A., Isik, S., & Kalin, L. (2014). Flood flow forecasting using ANN, ANFIS and regression models. *Neural Computing and Applications*, *25*, 25-37.

[33] Saha, A., Pal, S. C., Arabameri, A., Blaschke, T., Panahi, S., Chowdhuri, I., ... & Arora, A. (2021). Flood susceptibility assessment using novel ensemble of hyperpipes and support vector regression algorithms. *Water*, *13*(2), 241.

[34] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

[35] Wikipedia contributors. (2023). Coefficient of determination. *Wikipedia*. https://en.wikipedia.org/wiki/Coefficient_of_determination.

[36] Faraway, J. J. (2016). Does data splitting improve prediction? *Statistics and computing*, *26*, 49-60.

[37] Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531-538.

[38] Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., ... & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*, 1-15.

[39] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

[40] El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* (pp. 3-11). Springer International Publishing.

[41] Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, 234-240.

[42] Barde, V., Nageswararao, M. M., Mohanty, U. C., Panda, R. K., & Ramadas, M. (2020). Characteristics of southwest summer monsoon rainfall events over East India. *Theoretical and Applied Climatology*, *141*, 1511-1528.

[43] Wikipedia contributors. (2023). 2022 South Asian floods. *Wikipedia*. https://en.wikipedia.org/wiki/2022_South_Asian_floods.

[44] Davies, R. (2022). *Saudi Arabia – Severe Floods Hit Jeddah After 179mm of Rain in 6 Hours*. https://floodlist.com/asia/saudi-arabia-floods-jeddah-november-2022.

[45] Wang, J., & Chen, T. (2014). An online method to remove chattering and repeating alarms based on alarm durations and intervals. *Computers & Chemical Engineering*, *67*, 43-52.

[46] Brakenridge, G. R., Nghiem, S. V., Anderson, E., & Mic, R. (2007). Orbital microwave measurement of river discharge and ice status. *Water Resources Research*, *43*(4).

[47] Khalaf, M., Hussain, A. J., Al-Jumeily, D., Fergus, P., & Idowu, I. O. (2015). Advance flood detection and notification system based on sensor technology and machine learning algorithm. In *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 105-108). IEEE.

[48] Misra, P., & Yadav, A. S. (2019). Impact of preprocessing methods on healthcare predictions. In *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*.

[49] Wen, X., Kleinman, K., Gillman, M. W., Rifas-Shiman, S. L., & Taveras, E. M. (2012). Childhood body mass index trajectories: modeling, characterizing, pairwise correlations and socio-demographic predictors of trajectory characteristics. *BMC medical research methodology*, *12*, 1-13.

[50] Cutts, C. S., & Eglen, S. J. (2014). Detecting pairwise correlations in spike trains: an objective comparison of methods and application to the study of retinal waves. *Journal of Neuroscience*, *34*(43), 14288-14303.

[51] Khalili, M., Brissette, F., & Leconte, R. (2009). Stochastic multi-site generation of daily weather data. *Stochastic Environmental Research and Risk Assessment*, *23*, 837-849.

[52] Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal*, *60*(3), 431-449.

[53] Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and information technologies*, *19*(1), 3-26.

[54] Breiman, L., & Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(1), 3-54.

[55] Bakar, N. M. A., & Tahir, I. M. (2009). Applying multiple linear regression and neural network to predict bank performance. *International Business Research*, *2*(4), 176-183.

[56] Wikipedia contributors. (2023). Linear regression. *Wikipedia*. https://en.wikipedia.org/wiki/Linear_regression.

[57] Bermolen, P., & Rossi, D. (2009). Support vector regression for link load prediction. *Computer Networks*, *53*(2), 191-201.

[58] Zheng, S. (2015). A fast algorithm for training support vector regression via smoothed primal function minimization. *International Journal of Machine Learning and Cybernetics*, *6*, 155-166.

[59] Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural computation*, *12*(5), 1207-1245.

[60] Rathore, S. S., & Kumar, S. (2016). A decision tree regression based approach for the number of software faults prediction. *ACM SIGSOFT Software Engineering Notes*, *41*(1), 1-6.

[61] Erdebilli, B., & Devrim-İçtenbaş, B. (2022). Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey. *Mathematics*, *10*(14), 2466.

[62] Zhang, Y., Liu, J., & Shen, W. (2022). A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, *12*(17), 8654.

[63] Bashar, S. S., Miah, M. S., Karim, A. Z., & Al Mahmud, M. A. (2019). Extraction of heart rate from PPG Signal: a machine learning approach using decision tree regression algorithm. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). IEEE.

[64] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3-29.

[65] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804-818.

[66] Seiler, R. A., Hayes, M., & Bressan, L. (2002). Using the standardized precipitation index for flood risk monitoring. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *22*(11), 1365-1376.

[67] Khole, M., Sreejith, O. P., Pai, D. S., & Devi, S. (2021). PUNE-411 005 INDIA.