Turkish Journal of Computer and Mathematics Education Vol.12 No. 7 (2021), 1762-1776

Research Article

Prediction Of Rainfall With A Machine Learning Approach

[1]Veera Ankalu.Vuyyuru, [2]Giduturi.Apparao , [3]S. Anuradha

[1]Research Scholar, CSE Department, GITAM Deemed To be University,Visakhapatnam,AP,India.
[2]Professor,CSE Department,GITAM Deemed To be University,Visakhapatnam,AP,India.
[3]Assistant Professor,CSE Department,GITAM Deemed To be University,Visakhapatnam,AP,India,
[1]veeraankalu14@gmail.com, [2]apparao.giduturi@gitam.edu, [3]asesetti@gitam.edu

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract- Machine Learning (ML) is a versatile method for working with complicated structures. This thesis focuses on creating a machine learning-based system for forecasting rainfall. The proposed ensemble model is provided meteorological data correlated with rainfall variables as an input. We are gathering data from IMD for this work (Indian Meteorological department). Validation and verification was carried out using ensemble learning with Support Vector Regression and Random Forest model (SVR-RF). Validation is conducted using usable measures from a meteorological department over a certain area, which aids in predicting the likelihood of rainfall. As a result, a novel solution is built to improve the system's efficiency by combining the promise of SVR-RF with the accuracy of rainfall prediction. By calculating the precipitation characteristics, the recorded data is used to forecast rainfall with continuous observations over a given area. The proposed model aids in the establishment of a partnership between rainfall variables and other similar variables, which benefits the proposed SVR-RF model's potential. Mean Absolute Error (MAE), Root Mean Square (RMS), and classification precision (day and monthly basis) are the output metrics used in the simulation. The proposed model has the ability to outperform current prediction models. The proposed model predicts rainfall effectively using a variety of measures such as temperature, precipitation, and so on. This model increases machine efficiency thus lowering error rates.

Keywords: rainfall prediction; Support vector regression; random forest; root mean square error; classification accuracy

1. Introduction

Rainfall is an observable fact of the climate system that has a significant effect on water-based production, resource planning, and the ecological system in general [1]. When it comes to the financial section, the amount of rain that falls over a period of time is used to estimate financial security. Over the last few decades, researchers have focused on forecasting the essence of rainfall using a variety of approaches to improve rainfall prediction precision.

However, the time and effort placed into modeling modern strategies paves the way for effective rainfall forecasting [2]. Some weather concepts and certain general derivations are present in the derivations. It is represented by arrangements between two parties under which the benefit is based on financial capital relevant to prediction. As a result, tools are weather types defined as rainfall when producing weather derivatives [3]. The popular difference between weather and other derivatives is dependent on an unidentified contract price [4]. As a consequence, numerous investigators have suggested various approaches and derivatives that are not appropriate for efficient rainfall prediction [5]. Rainfall estimation can be exceedingly challenging owing to certain physical processes. Furthermore, rainfall forecasting is closer to resident existence [6]. In particular, heavy rain is a natural phenomenon that triggers water logging and forecasts water accumulation over time. Similarly, real-time rainfall prediction with higher precision is expected in a timely manner based on complex dynamical changes in the atmosphere [7]. In the world of meteorology, this is viewed as a major task.

Machine learning and atmospheric simulations are the two main tools for forecasting rainfall at the moment. Atmospheric activities are carried out by the atmospheric model. It is determined that the equation is a closed structure of atmospheric motion [8]. This equation is used to forecast environmental patterns (rainfall) as well as ambient physical quantities. This model is further subdivided into climatic models, air circulation models, and computational models dependent on the criteria [9]. The air circulation model is used to calculate physical amounts, while the climatic model is used to calculate climatic conditions such as long-term rainfall. In the same manner, the numerical model is the most effective model for forecasting short- and medium-term rainfall [10]. The overall prediction precision is dependent on rational approximation since this equation-based approach depends on partial differential equations.

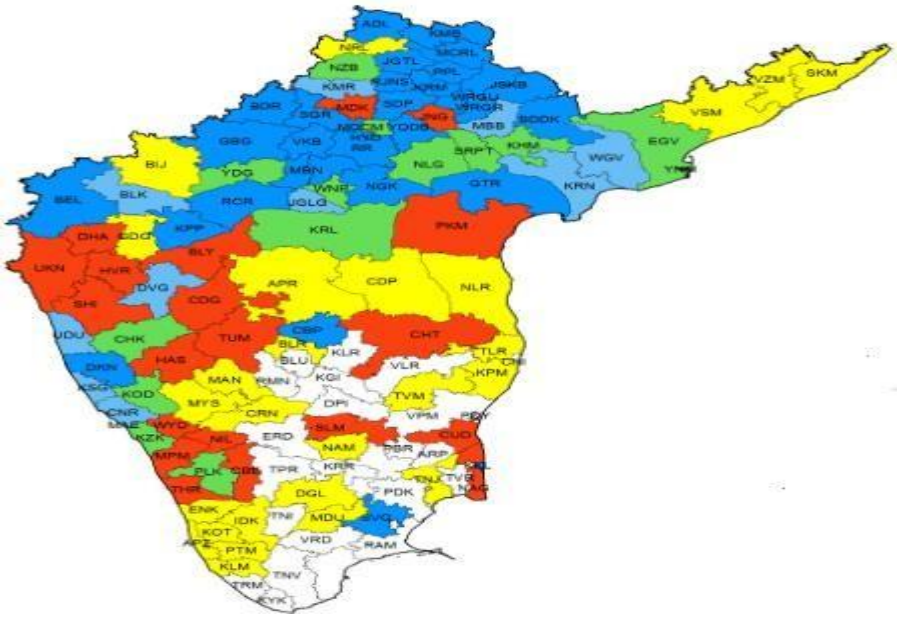

Fig 1. South Peninsular region rainfall conditions (Source: mausam.imd.gov.in)
Blue color- larger extent (60%), light blue – excess (20% to 50%), green – normal (-19% to 19%), red – deficient (-59% to -20%), yellow- large deficient (-99% to 69%), no color- no rain

Furthermore, the atmospheric model necessitates a significant amount of computing power. The conditions for forecasting rainfall are incredibly difficult to put into practice. As a result, using machine learning methods and intelligent device simulation, this research aims to address the above issues. This model is incredibly easy to grasp and trust when it comes to addressing different problems of rainfall forecasting. Machine learning techniques, on the other hand, provide a broad variety of applications for grappling with real-time problems. Under this definition, the rainfall prediction method is based on the discovery of a rule set or equation. It is intended to support investigators and is in charge of delivering correct reports.

The pitfalls that numerous researchers have experienced have caused them to switch to machine learning methods in order to provide an effective prediction model. The essence and dynamics of rainfall was captured and predicted in this simulation. The ultimate goal is to investigate the data by patterning or structuring it. As a consequence, a sophisticated paradigm to cope with this dilemma arises. Furthermore, one of the most significant drawbacks of machine learning methods is the high cost of tuning, as well as the difficulty of predicting the best model. As a result, the issue of mis-specification is solved by implementing a tuning protocol. The algorithm employs a more appropriate model, data representation with a broader variety of climates, and prediction based on new data. As a result, machine learning methods are more computationally efficient when compared to the equation type, which needs more tuning. As a consequence, this work highlights two main concepts.

contributions: 1) detailed study of cumulative rainfall on a regular basis; 2) data association assessment with validation using ensemble Help Vector Regression and Random Forest models (SVR-RF). By calculating the precipitation characteristics, the recorded data is used to forecast rainfall with continuous observations over a given area. For confirmation, the online accessible Meteorological dataset is used. The simulation is carried out in a Python environment. Accuracy, mean square error, and root mean square are the parameters used to evaluate the results. As opposed to other ones, the proposed model outperforms them.

Section 2 contains history analyses, section 3 contains methods, section 4 contains numerical findings with discussions, and section 5 contains the conclusion and prospective study directions.

2. Literature Survey

Statistical theory is used in standard Machine Learning (ML) methods. The time-series analysis is typically a multivariate statistical analysis. It's used to estimate rainfall dependent on time-series characteristics. Various

approaches are used in this study, including the Markov model, wavelet, and auto-regressive integrated moving average model (ARIMA). These methods begin with rainfall sequences and present a range of rainfall series characteristics. Furthermore, they do not focus on climatic influences on environmental conditions, instead focusing on rainfall directions that cause environmental shifts.

The following machine learning methods aid in the discovery of relationships between physical, meteorological, and rainfall variables. Support Vector Regression, Artificial Neural Networks, Support Vector Machines, Grey forecasting, and other methods are used to model rainfall. When compared to statistical models, these approaches can provide better prediction results. Various machine learning models have differentiating consequences when it comes to forecasting rainfall. Prediction and factor selection are two essential parameters in ML algorithms that provide a superior and resourceful prediction model. To be more precise, selecting suitable machine learning algorithms to carry out the predictions is critical. Radial Basis Function Neural Network (RBFN), Back Propagation Neural Network (BPNN), Feed-Forward Neural Network (FNN), and other ANN methods for rainfall prediction are commonly used.

In [11], Maity et al. use ANN to forecast rainfall in a variety of locations, including geographical regions. Two secret layers are used in the modeled network. Experiments show that the expected model outperforms the standard model. Similarly, the author used ANN to forecast rainfall in the chosen area on a seasonal and monthly basis. The network calculates temperature and the Nino index using climatic circulation and meteorological considerations as inputs. In addition, the author looked at ANN optimization with sufficient input variety. The author uses FFNN to forecast rainfall in the northern basin area, which has a lot of secret layers. In [12], Wu et al. use sea surface temperature and height-field as RBFNN and PCA input parameters, respectively. This model is used to forecast annual rainfall throughout the country's central area. The traditional NN has shallow layers and pretends to collapse inside local minima, preventing over-fitting problems.

In [13], the author proposes a rainfall forecast model based on DBN. The model's validity is demonstrated by extensive testing in two areas. Sankhadeep et al. [14] studied the interaction between spring rainfall and climatic conditions. On the basis of regression equations, he estimates spring rainfall. The precision and generalization efficiency of the model provides better results than the other models, according to the observations. Furthermore, this model focuses on a particular rainfall area. In [15], Haider et al. predict a novel solution based on a memory model. From radar photographs, which is used to determine cloud directions and structure. It is well established that short-term rainfall is finer and has a better forecast impact than traditional models. According to the findings of the aforementioned review, most methods also have weaknesses and struggle to address the technological challenges. They don't take generalization into account, and an aspect relevant to area selection isn't thoroughly validated. In the proposed model, this is provided a lot of thought.

3. Mechine Learning approaches

This segment explores the ensemble SVR-RF model that is used to forecast rainfall in India's southern area. Ensemble learning is a model for making decisions dependent on a variety of various models in general. The ensemble seems to be more resilient (no bias) and less data-sensitive by integrating/combining the individual models (less variance). The confirmation data was gathered from the Indian Meteorological Department [16]-[17]. Table I lists the variables in the model dataset.

a. Random Forest (RF)

Breiman developed the Random Forest (RF) model, which is commonly used in regression and classification research without the need to tune hyper-parameters. This technique is used to train a large number of DT predictors and then averaging them to minimize over-fitting and increase precision. Furthermore, non-linear interaction dynamics between the expected and predictor variables are well captured. This model will yield unbiased deviations (estimates) in classification and regression models for score estimation.

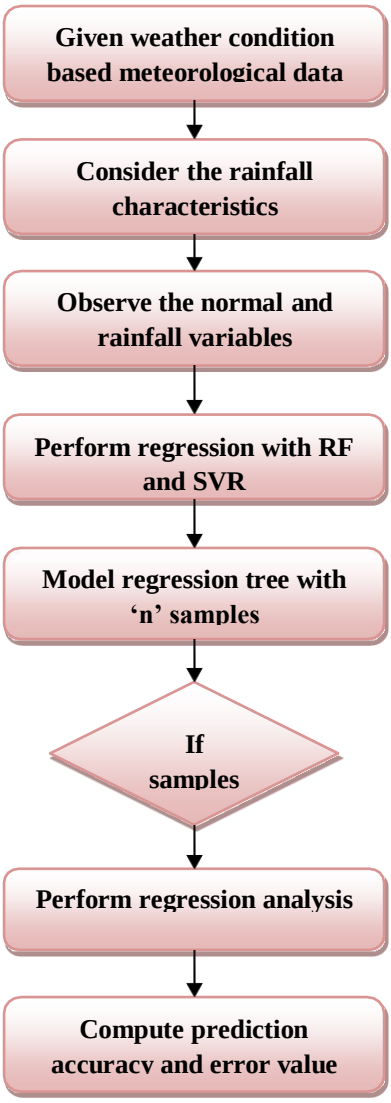

Fig 2: Flow diagram of the Ensemble RF-SVR model

The RF ensemble paradigm employs bagging as an ensemble system and a decision tree as an entity model. This model is ensemble-based in order to boost prediction precision and have a higher prediction score for the predictor variables used to train the random forest variables. Among the benefits described above, this model has certain interpretability and mathematical theory flaws, making it difficult to demonstrate the model's judgment.

Algorithm 1: Random Forest

Choose $'n'$ random subsets from the available training set
Train the given $'n'$ samples One random sub-set is utilized for training
Optimal split of all tree based on random subset features, i.e., assume 10 features and select 5 features randomly from the 10 features
The individual tree predicts candidates/records in test set independently
5. Make final prediction

From the above algorithm, it is known that ensemble classifier is extremely powerful which can be used for both regression and classification problem. Similarly, other individual models such as SVR can be applied with

boosting or bagging ensemble to acquire better performance.

Similarly, it lacks in predicting the values that goes beyond the training data ranges. This model randomly selects the values or samples (bootstrap) from diverse datasets (here, meteorological data) to model a decision tress. For every sample, the predicator variables subset is chosen randomly with least residual sum of squares to make the model works well. The forest (trees) has to ensure the growth to a larger extent. The derived final predictor values are averaged from the complete regression tree and used for computing the test dataset of the forests. With the tree growth, 1/3 of meteorological samples are provided for prediction estimation and the variable significance is utilized for further tree construction. For theoretical modeling, the method considers SVR for ensemble working principle to avoid the over-fitting issues.

b. Support Vector Regression (SVR)

The classification technique that relies over SVM for mapping the independent variables of available samples into space dimensions. Typically, it is utilized for categorizing the observations among two diverse groups. This is initially designed by Vapnik [18]. The author applies $\{(x_k, y_k)\}$ training observations for constructing the linear model with non-linear thresh old classifications by mapping variables on huge number of observations. The separation among the classes is attained with optimal hyperplane which is measured relies over $'N'$ observations. Here, $'x'$ is independent variable vector and $'y'$ is classification of all samples $\{-1, 1\}$. Therefore, the hyperplance classification is provided as in Eq. (1):

$$w^T \emptyset\,(x_k) + b = 0 \qquad (1)$$

The Eq. (1) above has to fulfill the conditions of Eq. (2) & (3):

$$w^T \emptyset\,(x_k) + b \geq 1; \qquad y_k = 1 \qquad (2)$$

$$w^T \emptyset\,(x_k) + b \leq -1; y_k = -1 \qquad (3)$$

This work considers the successive notations as $x \rightarrow$ predictor variables (vector); $y \rightarrow$ sample classification; $w \rightarrow$ weight vectors; $b \rightarrow$ constant; C, $c \rightarrow$ parameters.

The objective is not classified essentially into groups; however it estimates the real values. However, this model uses SVR to acquire regression model utilized to identify real-value measurements. The function maps the independent variables $\emptyset\,(.)$: $R^n \rightarrow R^{nk}$ towards the space with higher dimensionality where it is probable to linearly separate the samples based on the above Eq. (2), by fulfilling the conditions of Eq. (3) & (4):

$$y_k[w^T \emptyset(x_k) + b] \geq 1 \qquad (4)$$

The classification condition is $(x) = sgn\,(w^T \emptyset(x) + b)$. Moreover, the newer space is mapped with function $\emptyset(.)$; however, there is no proper separation $'N'$ into two classes $\{-1, +1\}$. Therefore, a variable $\xi \geq 0$ is defined as tolerance threshold margin in classification. The classification is flexible by accepting errors. With this flexible condition, the problem for determining the optimal hyperplane is convex optimization problem as in Eq. (5). With this equation, $'C'$ is a adjustment parameter of hyperplane edges with possible misclassification as in Eq. (6):

$$y_k[w^T \emptyset(x_k) + b] \geq 1 - \xi_k \qquad (5)$$

$$\min_{w,b} \frac{1}{2}\|w\| + C \sum_{k=1}^{N} \xi_k \qquad (6)$$

From the above Eq. (6), the optimization is converted to dual function with $y^T \propto = 0 e 0 \leq \propto_i \leq C; i = 1, \ldots, N$. Here, $e = [1, \ldots, N]^T$ is unit vector values, Q is $l * l$ matrix with $Q_{ij} \equiv y_i y_j K(x_i, x_j)$. This $(x_i, x_j) \equiv \emptyset\,(x)\emptyset(x_j)$ is known as kernel function.
The solution of the classification decision function is expressed as in Eq. (7):

$$sgn\,(w^T\emptyset(x) + b)$$
$$= sgn\,(\sum_{i=1}^{N} y_i \propto_i K(x_i, x) + b)) \quad (7)$$

From the above Eq. (7), it is known that SVR works alike of SVM, however the response variable is considered to be a continuous value $y \in R$. Here, SVR considers the linear regression function instead of considering the hyperplanes as in Eq. (8). The threshold error $\in$ is defined as minimized expression. This is termed as ε sensitivity loss error. This proposed SVR intended to reduce the ε error where $\|w\|^2$ is expressed as in Eq. (9):

$$f(x, w) = w^T x + b \quad (8)$$

$$|y - f(x, w)|_e = \begin{cases} 0 & |y - f(x,w)| \le \varepsilon \\ |y - f(x,w)| - \varepsilon & else \end{cases} \quad (9)$$

$$R = \frac{1}{2}\|w\|^2 + c\,(\sum_{i=1}^{N} |y_i - f(x_i, w)|\varepsilon) \quad (10)$$

The error tolerance variable is introduced and defined as ζ as excess value and $\zeta^* \ge 0$; and $i = 1, \ldots, N$.

$$\frac{1}{2}\|w\|^2 + c(\zeta_i + \zeta^*)_i \quad (11)$$

$$(w^T x_i + b) - y_i \le \varepsilon + \zeta_i \quad (12)$$

$$y_i - (w^T x_i + b) \le \varepsilon + \zeta^* \quad (13)$$

This work considers kernel function and provides linear, radial and polynomial function explicitly as in Eq. (14) – (16):

$$K(x_i, x_j) = x_i^T x_j \quad (14)$$

$$K(x_i, x_j) = e^{(-\gamma)\|x_i - x_j\|^2} \quad \gamma > 0 \quad (15)$$

$$K(x_i\, x_j) = (x_i^T x_j + 1)^d \quad (16)$$

The kernel function directly influences the values attained from SVR. The constants and parameters like $c, \gamma,$ are optimized. The training data is partitioned into two sets known as optimal parameters and validation of small error variables. This is termed as k-fold cross validation and chooses the above mentioned parameters based on lowest RMSE.

$$\min \frac{1}{2} \propto^T \propto$$

$$Q \propto\propto -e^T \propto$$

(17)

Algorithm 2: SVR

Input: available dataset = [X,Y]; X= input [array]; Y = c[array]; Output: prediction accuracy
Parameter initialization = variables ();
For learning the available variables do
Initialize velocity and position of i^{th} particle
For i = 1 do
Decode i^{th} particle
Determine SVR parameter (C,γ) and feature subset
Modify training set w.r.t. chosen feature subset
Establish SVR classifier using parameter values
Train SVR with training set
Compute test set with trained SVR
Determine overall fitness value
Optimize SVR parameters (C,γ) and feature subset
Establish SVR classifier with optimized parameter values
Evaluate test set with trained SVR
End for
End for
17. End

c. SVR-RF

Here, the powerful SVR and RF is ensemble to perform semi-supervised tasks. Initially, both these algorithms are ensemble to train the available labeled instances. However, SVR-RF is iteratively evaluated and finest outcomes are chosen to label the confident predictions. The process is chosen to label the confident prediction. This process is repeated till all the instances possess a label. Algorithm 3 gives the SVR-RF. The purpose of ensemble SVR and RF is given below. In general, RF is easier to tune when compared to SVR. It works faster than SVR based on the efficiency and amount of SVR data solver. It works effectually with categorical inputs where SVR needs to convert the categorical inputs into numerical form. Similarly, SVR works effectually with strange distance measures (kernel) than RF. Therefore, this ensemble method works better in various complex situations/environment. The flowchart of SVR-RF is given in Fig 3.

Algorithm 3: SVR-RF

Input: set of labeled and unlabeled instances

1. Initialize parameters, SVR classifier and RF classifier
2. Loop
3. While in unlabeled samples
4. Perform ensemble learning by splitting the training and testing
5. Compute the performance of SVR and RF after splitting
6. Choose classifier with higher accuracy score S_a
7. Train S_a by complete ensemble learning
8. Use S_a for selecting the confident predictions
9. Remove S_a from unlabeled instances
10. Add S_a to ensemble
11. Use S_a to predict the class labels of test cases

Here, hold-out method is utilized to select the appropriate model by eliminating the error drastically to train the data. The most-essential factor related to this ensemble is, the interpretable simplicity of prediction model. The training of anticipated model is performed in 15 iterations to label all the unlabeled samples. Hence, the numbers of instances labeled are dynamic in all iterations. In RF, 100 trees are constructed and in SVR, kernel radial bias function is used with $gamma$ = 32; $and\ C$ = 16respectively.

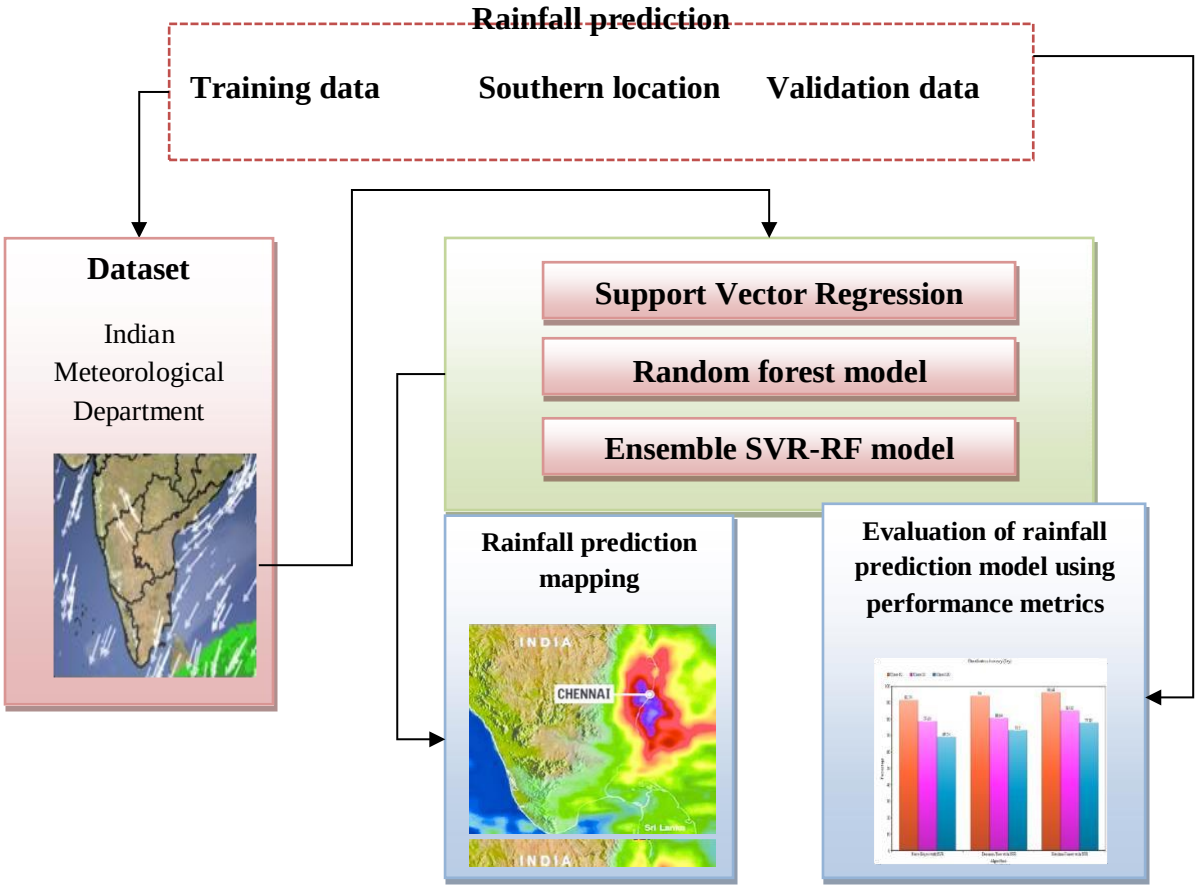

Fig 3: Architecture of SVR-RF

4. Numerical results

This investigation is extremely simple with the available toolkit. This is done with Python module and has been applied for executing the parameter tuning values, training data and prediction with RF algorithm. Typically, the

range of classification, clustering, and regression are merged with Python ML toolkit that includes RFs, and SVR. The prediction of rainfall is experimented with ensemble SVR-RF model. The outcomes are compared with random forest by considering the average returns of provided variance. The accuracy, correlation coefficient and the RMSE values are considered for computing the rainfall prediction in daily and monthly basis. The expression for RMSE is provided in Eq. (18):

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(d_i - \hat{d}_i)^2} \quad (18)$$

Here, $'T'$ are total test samples; $'d'$ are real sample values; $\hat{d}$ is estimated value. The correlation coefficient is linear association measure among the variables. The $'r'$ value is evaluated as in Eq. (19):

$$r = \frac{\sum_{t=1}^{n}(Y_t - \bar{Y})(Y' - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})\sum_{t=1}^{n}(Y' - \bar{Y})} \quad (19)$$

Here, Y_t and Y' is predicted/observed rainfall at $'t'$ time. $\bar{Y}$ is a mean of predicted/observed rainfall respectively. $'n'$ is total data points. $'r'$ Value lies among [-1, 1]. -1 is negative linear association; 1 is positive linear association; and 0 is no linear association. When $'r'$ value is higher, the model provides better performance. The R^2coefficient of SVR is compared to RF and Linear regressions. The values are 0.21, 0.14 and 0.0032 on daily basis. Similarly, the values are 0.71, 0.54 and 0.031 respectively. The MSE of SVR is 324.51 which is 10.11 and 16.73 lesser than RF and Linear regression. The RMSE values are 15.46, 16.32 and 18.25 for daily rainfall prediction. Similarly, the RMSE values are 168.51, 179.25 and 235.54 for monthly basis. Table II shows the error values of SVR, RF and linear model. Similarly, Fig 4 to Fig 7 depicts the rainfall graph of southern peninsular region during monsoon, pre- and post-monsoon and winter. Fig 8 to Fig 10 depicts the regression analysis of R^2 coefficients, MSE and RMSE values respectively.

Table I: Raw Data Collected from Indian Metrological unit

Sub-division	Year	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec	Annual
Andra Pradesh	1901	24.5	39.1	21.7	36	74	41.8	49	67.9	191	122	212	80	960
Andra Pradesh	1902	67.2	9.8	25.1	21.9	85	39.3	55	114	98.6	282	175	166	1138
Andra Pradesh	1903	19.3	7.8	1.7	18.2	129	58.5	73	115	210	128	201	203	1164
Andra Pradesh	1904	35.2	0.1	0.7	19.5	122	34.9	89	40.4	85.7	163	24	49	663
Andra Pradesh	1905	6.5	7.5	17.2	64.8	84	49.8	39	102	73.5	250	124	3.2	821
Andra Pradesh	1906	52.4	12.9	17	8.5	40	43.6	76	195	65.3	163	189	128	990
Andra Pradesh	1907	8.4	1.2	25	78.9	57	46.1	70	58.2	117	160	200	63	886
Andra Pradesh	1908	16.8	36.9	25	24.5	70	34.2	38	48.4	154	261	36	20	765
Andra Pradesh	1909	116.5	11.2	7.7	68.6	117	30.8	36	207	106	158	75	21	955
Andra Pradesh	1910	9.2	21.2	2.4	26.8	65	51.5	148	142	52.4	278	147	0.6	944
Andra Pradesh	1911	2.8	1.1	6	30.8	72	70.7	50	28.9	121	126	176	130	815
Andra Pradesh	1912	5.9	2.8	2.9	18.5	64	46.7	35	74.2	127	244	259	24	904
Andra Pradesh	1913	2.9	3.2	6.6	21	65	29.7	59	62.3	125	203	184	118	879
Andra Pradesh	1914	5.7	1.9	5.6	34.6	57	37.9	38	98.4	132	281	129	132	953
Andra Pradesh	1915	35.1	18.1	43.6	37	52	66.4	118	82.9	133	99.9	241	62	989
Andra Pradesh	1916	0.1	3.2	3.1	19.2	61	34.5	169	126	94.7	209	146	33	897
Andra Pradesh	1917	14.3	39.1	32.3	11.9	78	72.4	45	156	164	123	168	46	948
Andra Pradesh	1918	80.9	5.4	25.9	9.9	82	36.2	41	58.6	47	81.8	315	88	872

Andra Pradesh	1919	22.9	0.8	16.2	25	89	49.9	88	51.6	163	152	264	106	1029
Andra Pradesh	1920	135.9	3.1	3.7	67.3	60	47.3	37	74.8	156	159	405	4.7	1154
Andra Pradesh	1921	141.2	0	1.7	72.3	40	55.7	93	130	78.3	236	76	56	979
Andra Pradesh	1922	41.9	18.1	2.5	34.9	100	48.8	45	87.1	75.3	263	292	72	1081
Andra Pradesh	1923	136	2.5	50.1	26.6	31	33.7	41	39.8	99.8	220	44	110	834
Andra Pradesh	1924	29.7	0.7	31.3	30.2	75	54.9	114	87.1	171	107	164	31	895
Andra Pradesh	1925	17.6	4.6	57.2	30.1	85	50.6	39	91.7	77.3	149	229	183	1013
Andra Pradesh	1926	88	2.2	30.4	41.4	54	38.7	74	70.5	120	152	119	26	816

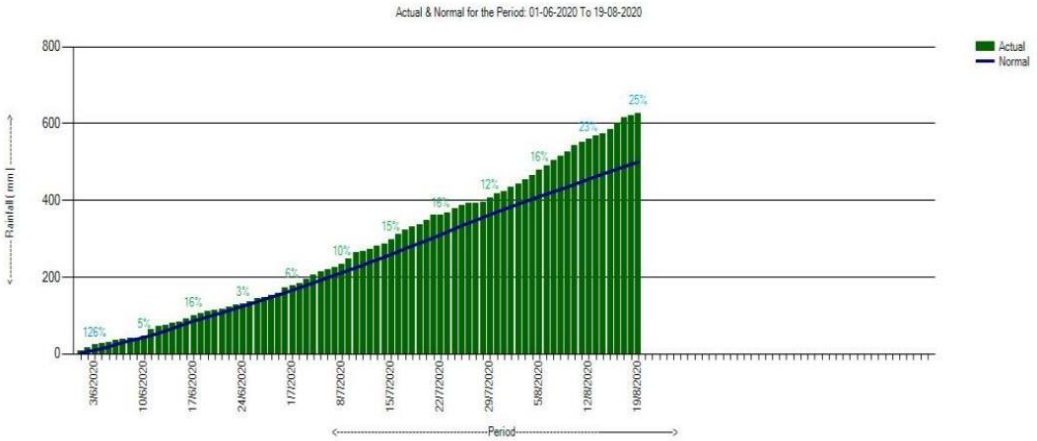

Fig 4: Rainfall graph in southern peninsular region [16] (During Monsoon)

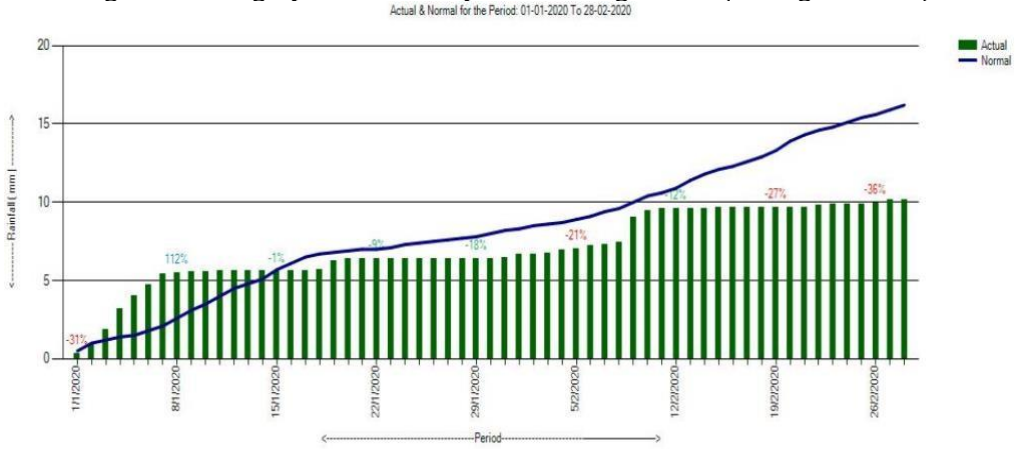

Fig 5: Rainfall graph in southern peninsular region [16] (During Winter)

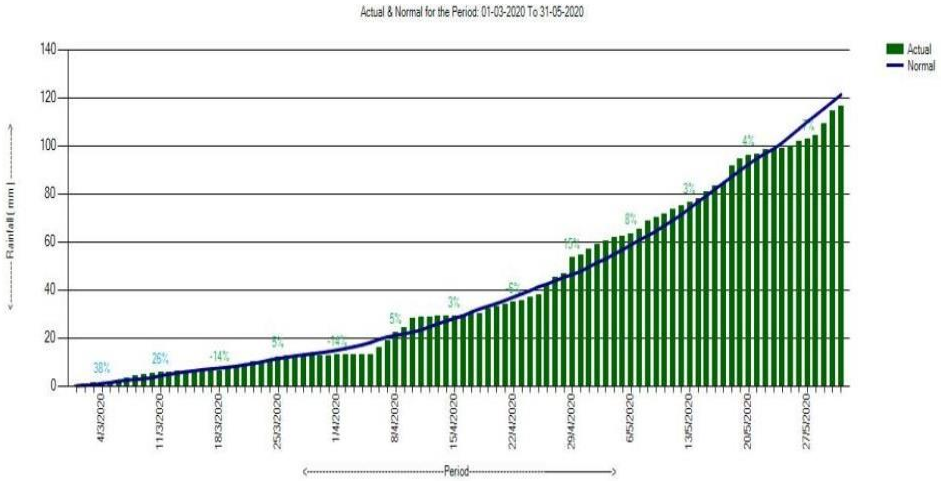

Fig 6: Rainfall graph in southern peninsular region [16] (During Pre-monsoon)

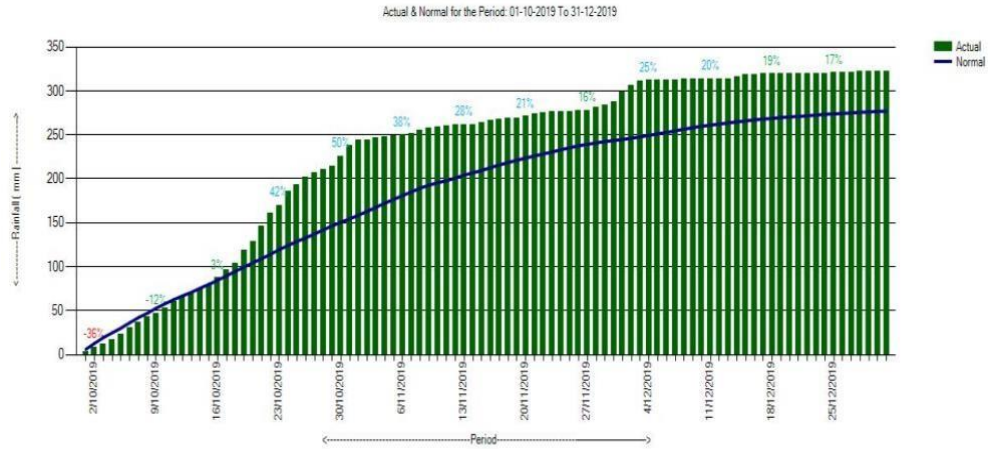

Fig 7: Rainfall graph in southern peninsular region [16] (During Post monsoon)

Table II: Error value computation

Regression Algorithm	R^2 Coefficient		MSE		RMSE	
	Day	Month	Day	Month	Day	Month
Support Vector	0.21	0.71	324.51	22678.32	15.46	168.51
Random Forest	0.14	0.54	334.62	32223.42	16.32	179.25
Linear	0.0032	0.031	341.24	47732.16	18.25	235.54

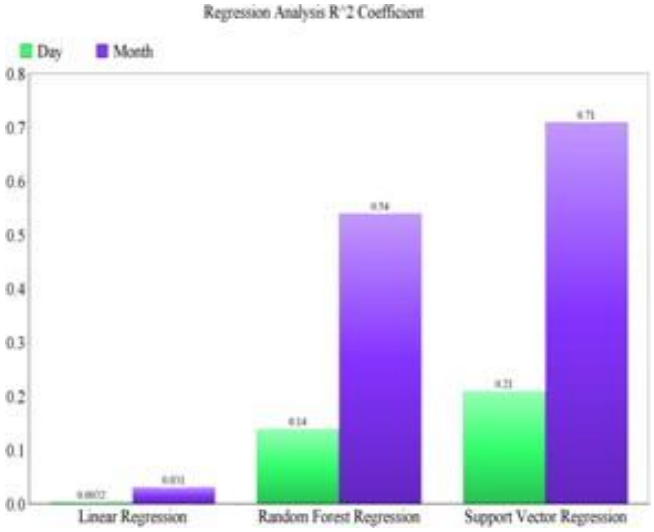

Fig 8: Regression analysis for R^2 coefficient

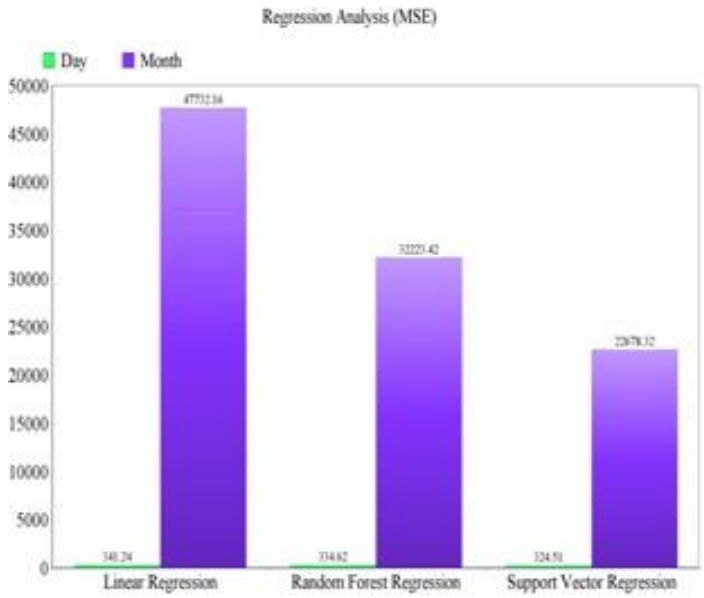

Fig 9: Regression analysis for *MSE*

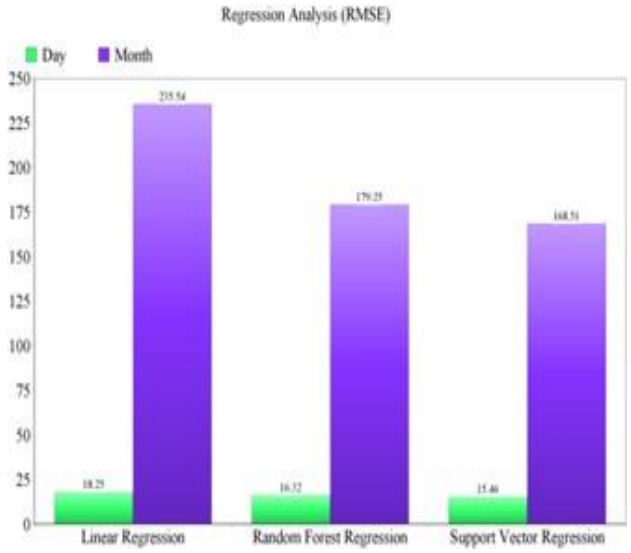

Fig 10: Regression analysis for *RMSE*

The SVR prediction results are compared with the prevailing approaches like linear regression and random forest regression. The data is observed on daily and monthly basis. The daily observations are extremely considered by the meteorological department. There are two diverse strategies are considered while using SVR. The former is to partition the data into training observations for SVR optimization and testing observations for predicting the errors. For partitioning, 70% is training and 30% is a testing data. This model is optimized dynamically and updated periodically. The longer period for data analysis is extremely significant and suitable for evaluating the prediction stability. The error rate of the SVR is compared with RF model on daily basis. SVR model shows linear kernel that has higher prediction power than that of RF. The error values of RF regression and linear regression is higher than SVR model respectively. The SVR error is fixed for training period.

Here, the classification accuracy of RF, DT and NB is compared. The accuracy of RF is 96.34% which is 2.34% lower than DT and 4.6% higher than NB respectively in daily basis. Similarly, the accuracy of RF is 74.32% which is 2.22% and 4.32% higher than DT and NB respectively in monthly basis. Table III depicts the classification accuracy of the anticipated RF with DT (Gini Index) and NB respectively. Here, three different classes are considered for computation, they are, class 10, class 50 and class 100 respectively. The graphical representation of prediction accuracy is given in Fig 11 and Fig 12 respectively.

Table III: Prediction accuracy

lassification Accuracy	**Classes**					
	10		**50**		**100**	
	Day (%)	**Month (%)**	**Day (%)**	**Mont h (%)**	**Day (%)**	**Month (%)**
SVR-RF	96.34	74.32	85.32	51.26	77.85	39.17
Decision Tree (Gini Index)	94	76.54	80.64	46	73.20	32
Naïve Bayes	91.74	70.26	78.64	41.24	69.24	29.54

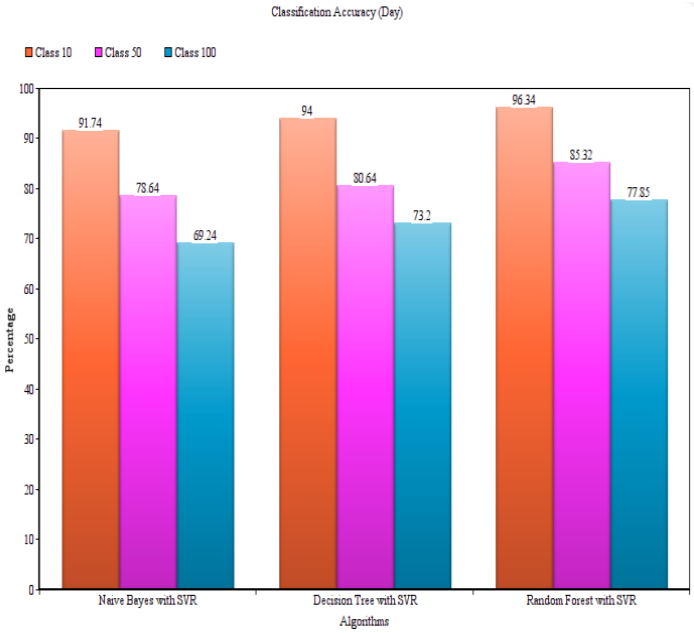

Fig 11: Classification accuracy (day basis)

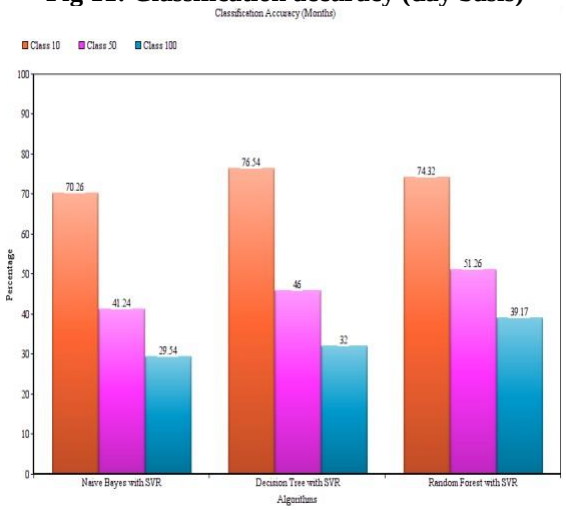

Fig 12: Classification accuracy (Month basis)

From the above analysis and tabular representation it is known that the anticipated SVR-RF works effectually in predicting the rainfall over the southern region. The error is drastically reduced with this regression model and an accuracy of 96.34% is attained with the daily observation condition. The analysis is done for monthly basis also where the accuracy is 74.32% which is higher than the other two models.

5. Conclusion

This work concentrates on providing the potential benefits of predicting rainfall using regression and ensemble the RF and SVR. This ensemble model is used for predicting rainfall in daily and monthly basis with the available meteorological dataset variables. The performance of the anticipated model is compared with existing NB, DT, LR and RF regression respectively based on the certain selected region for measuring the climate with rainfall characteristics. The ML approach has the ability to deal with the non-linear relationship among normal and the rainfall variables. Hence, it can be effectually applied to predict rainfall in prior state. The simulated variables for predicting rainfall provides beneficial results with 96.34% accuracy and RMSE of 15.46 respectively. The error rate of the anticipated model is lesser than the linear regression and RF regression and exploits its ability towards the prediction. This is extremely beneficial while considering the factors like observed data which is sparsely available over certain region. The performance of the anticipated model is higher than the prevailing approaches and gives better trade-off in contrary to other models. The simulation is done in Python environment. The daily/monthly rainfall magnitudes are noted with the proposed model which is better than the individual approaches. Therefore, the ensemble SVR-RF model is measured as a better choice for predicting the extreme rainfall prediction events.

In future, the prediction can be done with the deep learning approaches as the selections of features are

automatically done with the DL approaches. Also, the accuracy is expected to be higher than the anticipated model.

REFERENCES

1. M. Kühnlein, T. Appelhans, B. Thies, and T. Nauss, “Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI,” Remote Sens. Environ., vol. 141, pp. 129–143, Feb. 2014.
2. V. Bharti and C. Singh, “Evaluation of error in TRMM 3B42V precipitation estimates over the Himalayan region,” J. Geophys. Res., Atmos., vol. 120, no. 24, pp. 12458–12473, 2015.
3. Andrade José R., and Ricardo J. Bessa, "Improving renewable energy forecasting with a grid of numerical weather predictions," IEEE Transactions on Sustainable Energy, vol. 4, pp.1571-1580, 2017.
4. Niya Chen, Zheng Qian, Ian T. Nabney, and Xiaofeng Meng, “Wind power forecasts using Gaussian processes and numerical weather prediction,” IEEE Transactions on Power Systems, vol.29, pp.656-665, 2014.
5. Mao M., Ling J., Chang L., ET AL.: ‘A novel short-term wind speed prediction based on MFEC’, IEEE J. Emerg. Sel. Top.
6. Power Electron., 2016, PP, pp. 1–1
7. Cabrera, B. L. , Odening, M. , & Ritter, M. (2013). Pricing rainfall futures at the CME. Journal of Banking & Finance, 37 (11), 4286–4298 .
8. Cramer, S. , Kampouridis, M. , & Freitas, A. (2016a). A genetic decomposition al- gorithm for predicting rainfall within financial weather derivatives. In Gecco ’16: Proceedings of the 2016 genetic and evolutionary computation conference (pp. 885– 892). Denver, USA: AC
9. Cramer, S. , Kampouridis, M. , Freitas, A . A . , & Alexandridis, A . (2015). Predicting rain- fall in the context of rainfall derivatives using genetic programming. In 2015 ieee symposium series on computational intelligence (pp. 711–718).
10. Mislan , Haviluddin , Hardwinarto, S. , Sumaryono , & Aipassa, M. (2015). Rainfall monthly prediction based on artificial neural network: A case study in teng- garong station, east kalimantan - indonesia. Procedia Computer Science, 142–151. Int. conf on Computer Science and Computational In- telligence (ICCSCI 2015)
11. D. Kumar, A. Singh, P. Samui, and R. K. Jha, ``Forecasting monthly precipitation using sequential modelling," Hydrol. Sci. J., vol. 64, no. 6, pp. 690_700, Apr. 2019.
12. R. Maity, P. P. Bhagwat, and A. Bhatnagar, `”Potential of support vector regression for prediction of monthly stream_ow using endogenous property," Hydrol. Processes, vol. 24, no. 7, pp. 917_923, Mar. 2010.
13. Wu, K. W. Chau, and C. Fan, ``Prediction of rainfall time series using modular arti_cial neural networks coupled with data- preprocessing techniques," J. Hydrol., vol. 389, nos. 1_2, pp. 146_167, Jul. 2010.
14. J. Abbot and J. Marohasy, ``Input selection and optimization for monthly rainfall forecasting in Queensland, Australia, using arti_cial neural networks," Atmos. Res., vol. 138, pp. 166_178, Mar. 2014.
15. Sankhadeep Chatterjee, Bimal Datta, Soumya Sen, Nilanjan Dey, “Rainfall Prediction using Hybrid Neural Network Approach”, 2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom), pp. 67 – 72.
16. A. Haidar and B. Verma, ``Monthly rainfall forecasting using one dimensional deep convolution neural network," IEEE Access, vol. 6, pp. 69053_69063, 2018.
17. http://hydro.imd.gov.in/hydrometweb/
18. https://mausam.imd.gov.in/
19. Vapnik VN, “The Nature of Statistical Learning Theory”, New York: Springer; 1995.