Estimation of Rainfall Quantity using Hybrid Ensemble Regression

Preetham Ganesh
Department of Computer Science and Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Coimbatore, India
preetham.ganesh2015@gmail.com

Harsha Vardhini Vasu
Department of Computer Science and Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Coimbatore, India
harshavardhini2019@gmail.com

Dayanand Vinod*
Department of Computer Science and Engineering
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham
Coimbatore, India
v_dayanand@cb.amrita.edu

***Abstract*—Accurate prediction of rainfall in a geographical region has always been a challenge to the researchers. In this paper, ensemble methods such as bagging and boosting are used to predict rainfall level in districts belonging to Tamil Nadu, India. The Ensemble Regression models are optimised by tuning the parameters such as the number of estimators, base estimator and maximum depth. For evaluating the developed models, performance measures such as Mean Squared Error and Explained Variance Score were used. Based on the analysis, Bagging Regression produced better results than the other models after optimisation, but the difference between the performance of the models was very less. Hence, the prediction of the ensemble regression models is used instead of the features to predict rainfall, where two or more models are used at a time in different combinations for this purpose. The models are combined in different combinations using ensemble techniques such as Simple Averaging, Blending and Stacking. The developed models are compared using graphical analysis, where the comparison is based on actual rainfall values.**

***Index Terms*—Rainfall Prediction, Regression, Bagging, Boosting, Blending, Stacking, Hybrid Ensemble**

I. Introduction

Rainfall plays a vital role in the life of every living organism on this planet. It helps in maintaining the groundwater table on land, helps in balancing vegetation, i.e. without rainfall, trees would dry away, and the plants could die, leading to barren lands. In recent times, many cyclones occur at uneven periods, thereby destroying vegetation, killing animals and human beings and damaging public properties, which causes the government to spend huge amounts of money on the damages. Hence, predicting the rainfall can help the government and the public in many ways.

Generally, researchers have used regression or classification methods for predicting rainfall. However, for quantitative prediction regression should be used as it predicts a value rather than predicting a range to which the value belongs. The ultimate aim of every researcher who has worked in this field is to develop a model that can predict rainfall with low error, but they forget the fact that the model should be able to capture variation.

A standard tool used by machine learning researchers for prediction is Ensemble Learning methods. It uses multiple prediction methods to obtain better performance. However, the performance of the ensemble regression models can be enhanced to a greater extent by using Hybrid Ensemble Regression Models. A Hybrid Ensemble Regression Model is a combination of two or more ensemble regression models, where the models are combined using ensemble techniques.

To build Ensemble Regression Models methods such as Random Forest Regression (RFR), Extra Trees Regression (ETR), Bagging Regression (BAR), Gradient Boosting Regression (GBR) and Extreme Gradient Boosting Regression (XGBR) are used. The methods mentioned above are optimised to a great extent for obtaining better results. Likewise, for building hybrid ensemble regression models techniques such as Simple Averaging, Blending and Stacking were adopted.

The structure of the proposed work is as follows: Section II lists the previous works related to the rainfall prediction and the other applications of Hybrid Ensemble Methods. Section III explains the process flow for the proposed solution. Section IV discusses in detail about the derived results. Section V concludes the paper based on the derived results.

II. Related Works

This section discusses in detail about the previous works done by researchers in the prediction of rainfall and the other real-time applications where the ensemble methods are used. It also discusses the various papers in which hybrid ensemble techniques are used to optimise the results further.

The dataset used in this paper for analysis was downloaded from the India Water Portal - Met Data Repository and the papers using the same dataset are discussed below. S K Mohapatra, A. Upadhyay and C Gola in [1] used Multiple Linear Regression (MLR) to predict rainfall in the Bangalore District, Karnataka, India. The features used for prediction is Wet Day Frequency, and the prediction was made in a season-wise manner (Rainy, Summer and Winter). The authors compared the performance of validation techniques such as Holdout method and K-Fold Cross Validation method. Based on the analysis, it was concluded that for rainy and winter

978-1-7281-5523-4/19/$31.00

seasons, the K-Fold Cross Validation performed better and for the summer season, the Holdout Method performed better.

A H Manek and P K Singh in [2] used Back Propagation Neural Network (BPNN), Radial Basis Function Neural Network (RBFNN) and Generalised Regression Neural Network (GRNN) to predict rainfall in Nilgiris District, Tamil Nadu, India. The features used for prediction are Cloud Cover, Average Temperature and Vapour Pressure. All the neural networks mentioned above had a single hidden layer, and the architectures are 10-1, 10-1 and 90-1, respectively. The authors concluded that RBFNN performed better than the other models.

P Ganesh, H V Vasu and D Vinod in [3] used MLR, Polynomial Regression (PR), Decision Tree Regression (DTR) and Support Vector Regression (SVR) to predict rainfall quantity in all the districts of Tamil Nadu, India. Three models have been developed for this purpose, namely District-Specific model, Cluster-Based model and Generic-Regression model. The authors concluded that the Generic-Regression model performed better than the other models for most of the districts.

There are various datasets available for predicting rainfall in different regions across the world, and there are many papers using ensemble methods on those datasets, which are discussed below. C Valencia-Payan and J C Corrales in [4] used RFR, BAR, Stacking and Multiple Layer Perceptron (MLP) to predict rainfall using multiscale data obtained from the Tropical Rainfall Measuring Mission (TRMM) and the Geostationary Operational Environmental Science (GOES) program. The authors compared the performance of plain RFR and a Stacking based Hybrid Model combining RFR, BAR and MLP in different combinations. Based on the analysis, the authors concluded that the combined model of RFR and BAR performed better than the others.

Naive Bayes Classifier (NBC), Random Forest Classifier (RFC), Sequential Minimal Optimisation (SMO) and MLP was used by A K Sharma, S Chaurasia and D K Srivastava in [5] to predict rainfall in the selected districts of Uttarakhand, India namely Almora, Chamoli and Tehri Garhwal. The numerical values of rainfall are split into three thresholds, namely low, mid and high volume of rainfall. Based on the analysis, the authors concluded that RFC outperforms the other models for all the districts.

V P Tharun, R Prakash and S R Devi in [6] used SVR, RFR and DTR to predict rainfall in the Nilgiris District, Tamil Nadu, India. R^2 and Adjusted R^2 (a modified version of R^2, where the difference is the inclusion of the weekly predictor) scores are the performance measures used to evaluate the developed models. Based on the analysis, it was concluded by the authors that RFR performs better than SVR and DTR.

The authors in [7] used RFR, MLP, Classification and Regression Tree (C&RT), SVR and K-Nearest Neighbour Regression (KNN-R) to predict rainfall in the Oxford city, Iowa, United States. For this purpose, the authors used the above mentioned five algorithms to develop three models where the difference lies in the number of inputs, i.e. 36, 44 and 52, respectively. Based on the analysis, it was concluded that MLP outperformed the other algorithms in all the developed models.

SVR and the ensemble techniques such as Simple Average Ensemble, MSE Ensemble, Variance Weighed Ensemble are used by K Lu and L Wang in [8] to predict monthly rainfall mean in Guangxi, China. The dataset used has dates ranging from January 1965 to December 2009. On analysis, it was concluded by the authors that SVR performed better than the other models.

Researchers also use the ensemble methods for prediction in other real-time applications which are discussed below. AdaBoost.RT was used by R Priya and D Ramesh in [9] to predict the correct amount of Nitrogen-Phosphorous-Potassium (N-P-K) content in the soil. The updated AdaBoost.RT algorithm works on the threshold to differentiate between the correct and incorrect predictions. The authors concluded that the updated AdaBoost.RT performs better than the traditional AdaBoost for all the crop types and all the nutrient types. J Fan et al. in [10] used SVR and XGBR to predict daily global solar radiation using temperature and rainfall in humid subtropical climates of China. On analysis, it was concluded that XGBR was more stable and efficient than SVR.

Landslide prediction was performed by H Hong et al. in [11] using Decision Tree Classifier (DTC) as the base classifier in AdaBoost Classifier (ABC), Bagging Classifier (BAC) and Rotation Forest Classifier (RoFC) in Guangchang, Jiangxi, China. For this purpose, 237 locations were selected and for training and testing the models, and the locations were divided int 70:30 ratio. 10 Fold-Cross Validation was used to validate the models. On analysis, it was concluded that RoFC with DTC as the best model to predict landslides.

In [12] the authors used SVR, RFR, ETR and Regression Trees (RT) to predict solar thermal energy. The models developed are compared based on Root Mean Squared Error (RMSE) and Computational Cost. Based on RMSE value analysis it was inferred that RFR and ETR perform better than SVR and DTR, and based on Time-Complexity analysis it was inferred that DTR is the computationally most efficient method and SVR is the computationally least efficient method which three times higher than RFR and ETR.

RFR, GBR, XGBR and SVR has been used by A Torres-Barran, I Alonso and J R Dorronsoro in [13] to predict the wind energy and solar radiation. The main focus of authors was to prove that the ensemble methods such as RFR, GBR and XGBR performs better than the preliminary methods such SVR, and based on analysis it was concluded that GBR and XGBR predict wind energy in a broader geographical range, and RFR and XGBR predict solar radiation better than the other two.

The authors in [14] used DTC, K-Nearest Neighbour Classification (KNNC), Logistic Regression (LR), NBC, RFC, ABC and Support Vector Machine (SVM) for spam mail detection. For this purpose two different datasets were used, one with email headers and the other without email headers. The preliminary analysis on the datasets indicates that in both cases, SVM outperforms the other models. To enhance the

results, the authors used different types of representations such as Term Frequency and Inverse Document Frequency (TF-IDF), Singular Value Decomposition (SVD) and Non-Negative Matrix Factorisation (NMF). On analysis, it was concluded that TF-IDF and SVD with ABC outperformed all the other models in all the performance measures.

In [15], the authors used Bagging and Boosting on base classifiers SVM, NBC and Maximum Entropy Classifier (MEC) to predict the sentiment of tweets in Twitter Social Media Platform. For Selecting Features that influence the sentiment, methods such as Point Wise Mutual Information (PMI) and Chi-Squares were used. The authors concluded that ensemble methods produced better results that base classifiers. On a comparison between the ensemble methods, it was concluded that Bagging produced better results than the Boosting.

The results of ensemble models are enhanced by the use of ensemble techniques which are discussed below. H Liang, L Song and X Li in [16] used Lasso Regression, Elastic Net and RFR to build three single prediction models and combining these models using Stacking Ensemble to predict the rotation stress of Steam Turbine. For validating the models, 3-Fold Cross-Validation has been used. Based on the analysis, it was concluded by the authors that Stacking Ensemble performed better than the other models.

P. Disorntetiwat and C. H. Dagli in [17] used Simple Averaging (SA) on Generalised Regression Neural Network (GRNN) to predict financial forecasting. The developed model is trained on input discreetly using GRNN, and the SA is done on the multiple GRNN. For comparison purposes, two different datasets were used, namely the S&P 500 index and Current Exchange Rate. The authors concluded that using GRNN, along with SA, produced excellent results for both the datasets.

The authors in [18] used ensemble classifiers such as BAC, ABC and Stacking on base classifiers such as NBC, DTC, Rule Induction (JRip) and iBK (K-Nearest Neighbour Classifier) to predict intrusion detection in the dataset obtained from Lincoln Laboratory. 10-Fold Cross-validation was used to validate the developed models. It was concluded that ABC using DTC performs better than the others.

Among all the papers mentioned above, the papers which have both ensemble learning algorithms and preliminary machine learning algorithms, it can be observed that ensemble learning algorithms perform better than the latter. It can also be inferred that researchers have not done not much work in using ensemble learning algorithms to predict rainfall in a particular geographical region. Also, papers with ensemble techniques such as Simple Averaging, Blending and Stacking have results with these methods, as mentioned earlier, performing better than the plain ensemble models.

In this paper, we have used ensemble regression algorithms to predict rainfall in all the district belonging to Tamil Nadu, India. The ensemble regression algorithms are optimised based on empirical analysis, which is explained in a detailed manner in Section IV-B. Also, ensemble techniques such as Simple Averaging, Blending and Stacking are used to create Hybrid Ensemble Regression models to enhance the performance of the ensemble regression models. A comparison of actual rainfall values is performed between developed models using graphical representation. Also, the best model from the graphical analysis is compared with the other papers using the same dataset.

III. Process Flow

Fig. 1 shows the process flow for the proposed architecture.

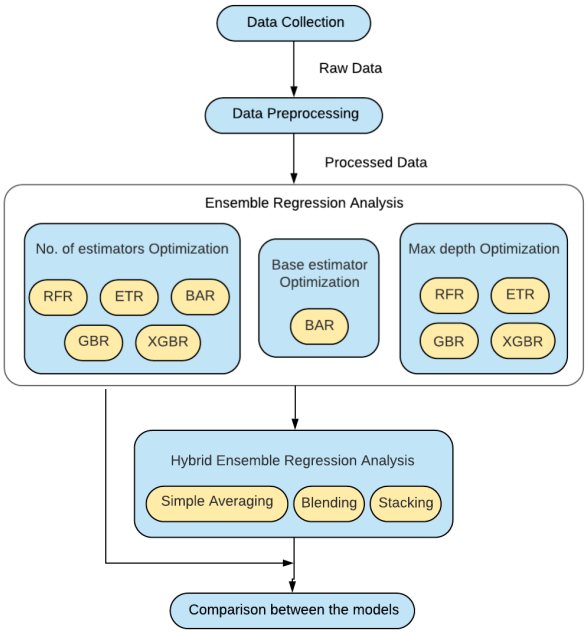

Fig. 1. Process Flow

A. Dataset Description

The dataset used for analysis is downloaded from the India Water Portal - Met Data Repository. The independent attributes in the dataset are 'Average Temperature', 'Maximum Temperature', 'Cloud Cover', 'Vapour Pressure', 'Crop Evapotranspiration', 'Potential Evapotranspiration' and 'Wet Day Frequency'. The dependent attribute is 'Rainfall'. The dataset, when downloaded was separate for each attribute for all the districts, i.e. each dataset consists of 12 columns and 102 rows, where each column is a month, and each row is a year. The period of the dataset ranges from 1901 to 2002.

B. Data Pre-Processing

The datasets of all features were combined to form one dataset consisting of 8 columns and 1224 rows for every district. All districts dataset are appended to form a combined dataset. The data was appended because the authors in [3] concluded that the regression algorithms modelled on the combined dataset performed best. A column named "District" (Not used for prediction) is added to the dataset which the district name to which tuple belongs. Since the columns in

the combined dataset have different ranges, the columns are normalised using Min-Max Normalisation. The formula to calculate the Min-Max Normalisation is given in (1).

$$A'_{\mathrm{i}} = \frac{A_{\mathrm{i}} - A_{\mathrm{min}}}{A_{\mathrm{max}} - A_{\mathrm{min}}} \tag{1}$$

Where A_i is the i^{th} element in the feature, A_{min} is the minimum value of the feature, A_{max} is the maximum value of the feature and A'_i is the normalised value of the i^{th} element in the feature.

C. Ensemble Techniques

1) Simple Averaging: Various prediction models are used to make predictions to a data point. The average of the predicted values of the models becomes the new predicted value.

2) Stacking: Stacking is also an ensemble learning technique, but it not only uses the testing set predictions but also uses training set predictions. The prediction made on the training set by multiple models is used as input to a regression algorithm to predict the actual dependent attribute.

3) Blending: It is similar to stacking, but the difference is that it uses the Holdout method instead of K-Fold Cross Validation method. The predictions are made on the holdout set, and the predictions of it are used to build a model executed on the test set.

D. Advanced Ensemble Algorithms

1) Bagging Regression (BAR): Bootstrap Aggregating (Bagging) is an ensemble learning algorithm which is designed to improve variance and avoid over-fitting. The base-estimator can be changed for better results. It is a modified case of the Model Averaging method. The model is created by using the sample with replacement technique on the original dataset where the ensemble model is a combination of individual prediction models. The model is trained on the sample dataset. The test set values are tested on the trained model, where the predicted value is the mean of all the predicted values in the individual models as given in (2) [19].

$$Y_{\mathrm{p}} = \frac{1}{N} \sum_{i=1}^{N} F_{\mathrm{i}}(X) \tag{2}$$

Where Y_p is the final predicted value of the model, N is the number of individual models in the ensemble, $F_i(X)$ is the predicted value of the individual model.

2) Random Forest Regression (RFR): Also known as Random Decision Forests is an ensemble learning method to perform regression and classification tasks. It constructs multiple decision trees during the training period and outputs the mean prediction of the individual trees (an extension of the bagging regression) [20]. The main difference is that at each feature split, arbitrary set of features are used. This process is also called as Feature Bagging.

3) Extra Trees Regression (ETR): It fits many completely extremely randomised decision trees (Extra Trees) on multiple samples of the dataset and averages the result to decrease the predictive error and controls over-fitting [21]. It differs from Random Forest by the following ways: (1) Each tree is trained on the entire training dataset compared to the Bootstrap in the RFR. (2) In trainer, the top-down splitting is arbitrary.

4) Gradient Boosting Regression (GBR): It produces an ensemble weak prediction models in a hierarchy fashion like the other boosting methods. It generalises them by allowing the modification of the arbitrary differentiable loss [22]. In a list of base models, it is assumed by the GBR that there is a faulty model among the individual models. GBR improves the faulty model by appending the estimator value to improve the prediction.

5) Extreme Gradient Boosting (XGBR): It is the advanced implementation of Gradient Boosting Regression which includes a variety of regularisation techniques and reduces overfitting and increases the performance [23].

E. Performance Measures

1) Mean Squared Error: It evaluates the excellence of a model where the error value is the mean squared difference between the actual and the estimated values in a list of prediction as in (3).

$$MSE = \frac{\sum_{i=1}^{N} (Y_{\mathrm{t}} - Y_{\mathrm{p}})^2}{N} \tag{3}$$

Where Y_t is the actual value, Y_p is the predicted value and N is the number of observations in the dataset.

2) Explained Variance Score: It is used to measure the discrepancy between a model and actual data. It is the part of the model's total variance that is explained by factors that are actually present and isn't due to error variance.

$$EVS = 1 - \frac{Variance(Y_{\mathrm{t}} - Y_{\mathrm{p}})}{Variance(Y_{\mathrm{t}})} \tag{4}$$

IV. Results and Discussion

The details of the results obtained on using various ensemble regression algorithms and hybrid ensemble techniques on modeling the rainfall data of districts belonging to Tamil Nadu, India is discussed in this section. The ensemble regression models are developed with the help of Sci-kit learn in Python and Hybrid Ensemble Regression models are developed in Python by the authors.

A. Parameters chosen for Optimisation of Ensemble Regression Algorithms

Tuning the parameters in the ensemble regression algorithms produces better results as the model will be able to predict rainfall with less error. The parameters tuned for the optimised fitting of the dataset for BAR are n_estimators and base_estimator. Likewise, for RFR, ETR, GBR and XGBR are n_estimators, max_depth, min_samples_split (The minimum number of samples required to split an internal node) and min_samples_leaf (The minimum number of samples required

to be at the leaf node). However, the performance analysis of min_samples_split and min_samples_leaf are not included in the paper because their tuning did not produce a noticeable difference in the performance measures.

B. Performance Analysis of the Ensemble Regression Models on the Generic Data

1) Number of Estimators (NoE) Optimisation: Number of Estimators (number of trees or models in the ensemble) plays a crucial role in the performance of an ensemble model. No matter how much the basic machine learning algorithms are tuned, the learning of information from the data fails beyond a point.

However, with ensemble models, increasing the number of estimators or models tends to boost the results. But, it loses its influence after increasing beyond a point because the learning becomes stagnant, and the model will not be able to learn anything new out of the data. For validating the developed models, K-Fold Cross Validation has been used where the number of folds is 10, and the number of repeats is 10. The MSE and EVS values for the RFR with the corresponding number of estimators are shown in Fig. 2 and Fig. 3.

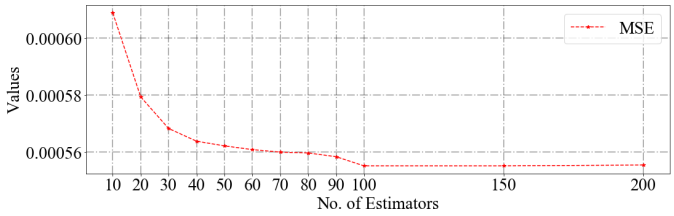

Fig. 2. RFR with different Number of Estimators versus their corresponding MSE values

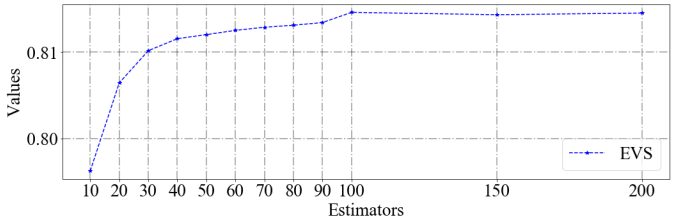

Fig. 3. RFR with different Number of Estimators versus their corresponding EVS values

It can be observed that as the number of estimators increases the MSE values tends to decrease in Fig. 2 and the R^2 values in Fig. 3 tend to increase, and it can also be observed that the figures are mirror images of one and another. The MSE and EVS values tend to become stagnant after estimators count reaches 100, so it can be concluded that the optimal number of estimators for RFR is 100 for the chosen dataset. The process mentioned above for RFR is extended to the other models, namely ETR, BAR, GBR and XGBR, and the optimal number of estimators has been identified and given in Table I.

Table I shows that for the current analysis RFR with the number of estimators equal to 100 performs better than the other models where the other parameters have the default values for all the models.

TABLE I
OPTIMAL NUMBER OF ESTIMATORS FOR ENSEMBLE REGRESSION MODELS

Ensemble Model	NoE	MSE	EVS
RFR	**100**	**0.000555**	**0.815**
ETR	90	0.000589	0.803
BAR	70	0.000559	0.813
GBR	50	0.000566	0.811
XGBR	50	0.000560	0.813

2) Maximum Depth Optimisation for RFR, ETR, GBR and XGBR: The parameter max_depth (Maximum Depth of the Tree) is only available for RFR, ETR, GBR and XGBR as these ensembles are based on Decision Trees. The MSE and EVS analysis of RFR for maximum depth ranging from 2 to 15 is given Fig. 4 and Fig. 5 for which the respective number of estimators is chosen from Table I.

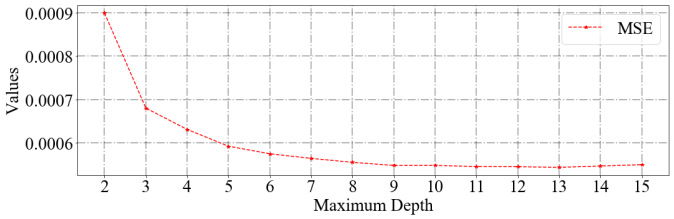

Fig. 4. RFR with different Maximum Depth versus their corresponding MSE values

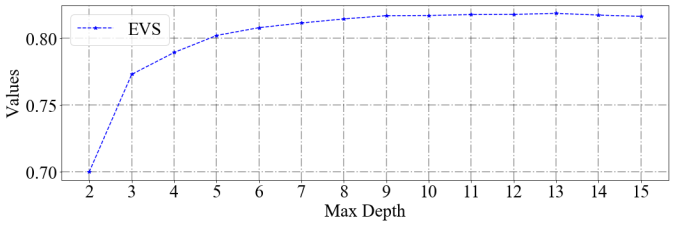

Fig. 5. RFR with different Maximum Depth versus their corresponding EVS values

Similar to the number of estimators analysis, it can be inferred from Fig. 4 and Fig. 5 that as the maximum depth increases the learning rate of the model tends to decreases. It can also be observed that learning stops after maximum depth value equal to 9. Hence it can be concluded that for RFR with the number of estimators equal to 100 the optimal Maximum Depth is 9. The process used to find the optimal maximum depth for RFR has also been used for other models, except for BAR (as it does not have the maximum depth parameter) for which the results are shown in Table II.

On observing the values in Table II, it can be inferred that the values of the performance measures are better than the values in Table I. So optimising the maximum depth has a reasonable impact on predicting rainfall. It can also be

TABLE II
OPTIMAL MAXIMUM DEPTH FOR ENSEMBLE REGRESSION MODELS

Ensemble Model	Maximum Depth	MSE	EVS
RFR	9	0.000548	0.8168
ETR	11	0.000546	0.8178
GBR	**5**	**0.000539**	**0.8199**
XGBR	5	0.000545	0.8190

observed that the best model in the maximum depth analysis is GBR with the number of estimators as 50 and the maximum depth as 5.

3) Base Estimator Optimisation for BAR: It is the base learner on which the boosting or bagging ensemble is constructed. For different base estimators, different performances of the bagging or boosting ensemble can be obtained.

In [3], for Generic-Regression model, four regression models namely Polynomial Regression with Degree equal to four (PR[4]), Decision Tree Regression with Maximum Depth equal to six (DTR[6]), Multiple Linear Regression (MLR) and Support Vector Regression with Kernel equal to Linear (SVR[L]) performed better than other models with different parameters. The above mentioned four models are chosen as base estimators for BAR. The analysis of it is given in Table III where the corresponding number of estimators is taken from Table I.

TABLE III
PERFORMANCE OF DIFFERENT BASE LEARNERS IN BAR

Base Learner	MSE	EVS
MLR	0.000646	0.784
DTR[6]	0.000576	0.808
PR[4]	**0.000517**	**0.827**
SVR[L]	0.001447	0.697

Similar to the result in our previous work, the values in Table III also indicate that BAR with PR[4] performs better than BAR with other base estimators. It can also be inferred that the same not only has better error values but also has better EVS value, indicating that the model is capturing variation better. Also, the MSE and EVS values of the model are better than GBR model in Table I. Also, the MSE and EVS values produced in Table III are better than the best model values in [3].

C. Prediction of Rainfall using Hybrid Ensemble Regression Models on the Generic Data

A Hybrid Ensemble Model is a combination of two or more different ensemble models, combined to improve the performance of the same. Our initial intention was to find the best ensemble regression model, but the performance measures of all the optimised models were almost similar, which led us to build a hybrid ensemble regression model.

1) Simple Averaging: The predicted values of the ensemble regression models from Section IV-B are taken, and the average of those values is computed, which is the new predicted value. The ensemble regression models are taken in different combinations for this purpose, to find the best combination that predicts rainfall the best.

2) Blending: A model based on blending is constructed by partitioning the entire dataset into train set and test. Since the dataset used is a combined dataset of all the districts belonging to Tamil Nadu, India, the last 224 records from each district is removed and attached to the test set leaving first 1000 records from each district for training. Hence the training set would have 29000 records (1000 records * 29 districts) and the test set would have 6496 records (224 records * 29 districts). If the process mentioned above was not followed, then the model will be trained on a few districts' data and will be tested on the remaining districts causing a significant error. The pictorial representation of the process used further for prediction is shown in Fig. 6.

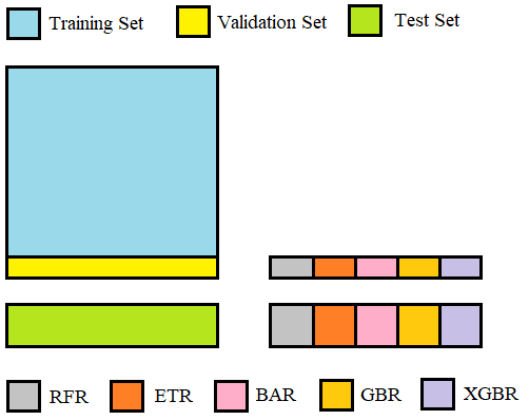

Fig. 6. Pictorial Representation of Blending based prediction

The training set is then split using the holdout method, where the validation set size is 0.1, and the records are picked randomly. All the ensemble regression models in Section IV-B are fitted on 90% of the training set, and the predictions are made for the remaining 10% (Validation set). Similarly, all the models are trained with the entire training set, and predictions are made for the test set. The stored predicted values of the validation set and test set respectively are used as the new training set, and the new test set where the models used are taken from Table III.

3) Stacking (K-Fold Cross Validation): The training and test set created for blending has also been used for stacking. However, unlike blending stacking uses K-Fold Cross Validation for creating the new training set instead of the holdout method. For every iteration, the prediction of the validation set has been stored and compiled together to form the new training set, as shown in Fig. 7. In this case, we have assigned K = 10. So, the new training set for stacking created is ten times larger than the training set created for blending.

Similar to blending the new test set has been created, and the newly created training and test sets are used for final predictions using the models from Table III.

TABLE IV
COMPARISON BETWEEN THE HYBRID ENSEMBLE REGRESSION MODELS USING MSE AS THE PERFORMANCE MEASURE

Combinations	**Simple Averaging**	**Blending**				**Stacking (Repeat=1)**				**Stacking (Repeat=10)**			
		MLR	**PR[4]**	**DTR[6]**	**SVR[L]**	**MLR**	**PR[4]**	**DTR[6]**	**SVR[L]**	**MLR**	**PR[4]**	**DTR[6]**	**SVR[L]**
RFR	0.000544	0.000333	0.000362	0.00048	0.003376	0.000323	0.000314	0.000343	0.0023	0.000323	0.000313	0.000325	0.00209
ETR	0.000556	0.00036	0.000342	0.000463	0.004041	0.000339	0.000323	0.000383	0.002242	0.000337	0.000321	0.000361	0.002385
GBR	0.000539	0.000332	0.000309	0.000456	0.003281	0.000311	0.000298	0.000366	0.002329	0.000312	0.000299	0.000305	0.002213
XGBR	0.000544	0.000319	**0.000298**	0.00044	0.003847	0.000296	**0.000291**	0.000354	0.00232	0.000297	0.000291	0.000301	0.002139
BAR	0.000518	0.00035	0.000499	0.000408	0.002806	0.000344	0.000339	0.000348	0.002	0.000336	0.000332	0.000337	0.001879
RFR, ETR	0.000545	0.000339	0.012075	0.000497	0.003419	0.000322	0.000328	0.000351	0.002042	0.000321	0.000324	0.00033	0.002103
RFR, GBR	0.000536	0.000328	0.003623	0.000483	0.002968	0.000313	0.000316	0.000353	0.002106	0.00031	0.000299	0.00031	0.001999
RFR, XGBR	0.000537	0.000324	0.010304	0.000492	0.003127	0.000301	0.0004	0.00038	0.002219	0.000302	0.000296	0.000316	0.002042
RFR, BAR	0.000515	0.000338	0.010862	0.000426	**0.002011**	0.000306	0.000606	0.000323	0.00183	0.000303	0.00028	0.000316	0.001647
ETR, GBR	0.000538	0.000329	0.000497	0.000485	0.003511	0.000313	0.000308	0.000358	0.002103	0.000309	0.000303	0.000315	0.002219
ETR, XGBR	0.000539	**0.000318**	0.001029	0.000461	0.00343	0.000299	0.000381	0.000338	0.002166	**0.000297**	0.0003	0.000316	0.002242
ETR, BAR	0.000519	0.000357	0.036623	0.000411	0.002397	0.00031	0.000405	0.000326	0.001815	0.00031	**0.000274**	0.000307	0.001846
GBR, XGBR	0.000538	0.000324	0.035758	0.000454	0.003269	0.000296	0.000292	0.000357	0.002217	0.000301	0.00029	**0.000301**	0.002212
GBR, BAR	**0.000513**	0.00036	0.063098	0.000516	0.00268	0.000309	0.000498	0.000319	0.001954	0.000302	0.00028	0.00031	0.001755
XGBR, BAR	0.000515	0.000355	0.004173	0.000431	0.002987	0.000298	0.000835	0.000331	0.001893	0.000297	0.000298	0.000308	0.001763
RFR, ETR, GBR	0.000538	0.000333	0.313515	0.000474	0.00309	0.000313	0.000389	0.000349	0.001945	0.000309	0.000309	0.000309	0.001968
RFR, ETR, XGBR	0.000538	0.000329	0.46725	0.000486	0.003181	0.000301	0.001212	0.000375	0.00206	0.000301	0.000325	0.000329	0.001987
RFR, ETR, BAR	0.000522	0.000341	0.14691	0.000438	0.002098	0.000305	0.001942	0.000315	0.001886	0.000303	0.000288	0.000305	0.001691
RFR, GBR, XGBR	0.000536	0.000323	0.217163	0.000506	0.003105	0.000301	0.000537	0.000381	0.002233	0.000304	0.000291	0.000315	0.001965
RFR, GBR, BAR	0.000518	0.000349	0.061342	0.000431	0.002449	0.000305	0.000531	0.000329	0.001839	0.0003	0.000293	0.000312	0.001733
RFR, XGBR, BAR	0.000518	0.000347	0.136336	0.000444	0.002642	0.000298	0.002017	0.000325	**0.001738**	0.000297	0.000293	0.000309	0.001706
ETR, GBR, XGBR	0.000536	0.000321	1.152639	0.000478	0.003228	0.000297	0.000534	0.000353	0.00223	0.000301	0.000295	0.000315	0.002166
ETR, GBR, BAR	0.000519	0.000361	0.287121	0.000491	0.002552	0.000305	0.000713	**0.000309**	0.001844	0.000301	0.000274	0.000303	0.001798
ETR, XGBR, BAR	0.000520	0.000356	0.430401	**0.000393**	0.002833	0.000297	0.002173	0.000331	0.001864	0.000297	0.000283	0.000304	0.001722
GBR, XGBR, BAR	0.000518	0.000358	0.239132	0.000444	0.002996	0.000297	0.014953	0.000318	0.001877	0.0003	0.000318	0.000306	0.001762
RFR, ETR, GBR, XGBR	0.000536	0.000329	2.621982	0.000525	0.003203	0.000299	0.006256	0.000376	0.002069	0.000304	0.000305	0.000327	0.001947
RFR, ETR, GBR, BAR	0.000522	0.000349	5.742796	0.000431	0.002429	0.000304	0.00074	0.000317	0.001823	0.0003	0.000333	0.000303	0.001692
RFR, ETR, XGBR, BAR	0.000523	0.000346	6.947006	0.000436	0.002672	0.000297	0.006977	0.000314	0.001745	0.000297	0.000291	0.000305	**0.001626**
RFR, GBR, XGBR, BAR	0.000521	0.000348	1.194721	0.000439	0.002715	0.000296	0.013717	0.000331	0.001945	0.0003	0.000342	0.000305	0.001723
ETR, GBR, XGBR, BAR	0.000521	0.000358	18.432151	0.000414	0.002931	0.000295	0.008582	0.00031	0.001857	0.0003	0.000353	0.000303	0.001797
RFR, ETR, GBR, XGBR, BAR	0.000524	0.000347	33.85816	0.00044	0.002756	**0.000295**	0.066363	0.000318	0.002014	0.0003	0.000369	0.000303	0.001692

TABLE V
COMPARISON BETWEEN THE HYBRID ENSEMBLE REGRESSION MODELS USING EVS AS THE PERFORMANCE MEASURE

Combinations	Simple Averaging	Blending				Stacking (Repeat=1)				Stacking (Repeat=10)			
		MLR	PR[4]	DTR[6]	SVR[L]	MLR	PR[4]	DTR[6]	SVR[L]	MLR	PR[4]	DTR[6]	SVR[L]
RFR	0.8182	0.8924	0.8827	0.8445	0.7257	0.8954	0.8985	0.889	0.8365	0.8956	0.8989	0.8949	0.8432
ETR	0.8143	0.8835	0.8892	0.8501	0.6731	0.8903	0.8956	0.876	0.8332	0.8909	0.8961	0.8831	0.8308
GBR	0.8199	0.8924	0.9	0.8523	0.7258	0.8995	0.9036	0.8817	0.8404	0.899	0.9031	0.9013	0.8397
XGBR	0.8202	0.8968	**0.9036**	0.8575	0.6952	0.9042	**0.9059**	0.8857	0.847	0.9039	0.9058	0.9026	0.8455
BAR	0.827	0.8868	0.8384	0.8681	0.7768	0.8887	0.8902	0.8873	0.8439	0.8912	0.8926	0.8909	0.8488
RFR, ETR	0.8178	0.8904	-2.9083	0.839	0.7475	0.8959	0.894	0.8864	0.8448	0.8963	0.8952	0.8933	0.8428
RFR, GBR	0.8208	0.8939	-0.172	0.8435	0.7659	0.8988	0.8978	0.8858	0.8469	0.8999	0.9033	0.8998	0.8494
RFR, XGBR	0.8209	0.8951	-2.3326	0.8406	0.7463	0.9025	0.8707	0.8771	0.8462	0.9023	0.9043	0.8979	0.8483
RFR, BAR	0.8278	0.8905	-2.5136	0.8622	**0.8212**	0.9008	0.8038	0.8953	0.8578	0.9018	0.9094	0.8978	**0.8642**
ETR, GBR	0.8201	0.8935	0.8391	0.8431	0.7152	0.8988	0.9003	0.8841	0.8483	0.9	0.9021	0.8979	0.8431
ETR, XGBR	0.8202	**0.8972**	0.667	0.8507	0.7218	0.9034	0.8767	0.8907	0.8486	**0.9041**	0.9029	0.8979	0.8453
ETR, BAR	0.8264	0.8845	-10.8559	0.8668	0.8018	0.8997	0.869	0.8945	0.8554	0.8995	**0.9113**	0.9005	0.8533
GBR, XGBR	0.8206	0.8952	-10.5741	0.8529	0.7254	0.9042	0.9058	0.8845	0.8492	0.9026	0.9063	**0.9027**	0.8426
GBR, BAR	**0.8285**	0.8834	-19.4196	0.8331	0.7781	0.9	0.8388	0.8969	0.8525	0.9023	0.9095	0.8997	0.8599
XGBR, BAR	0.8285	0.8851	-0.3508	0.8604	0.7595	0.9035	0.7297	0.8927	0.8598	0.9037	0.9034	0.9003	0.8603
RFR, ETR, GBR	0.8203	0.8922	-100.4767	0.8465	0.7494	0.8988	0.8741	0.8871	0.8517	0.9001	0.9	0.9001	0.8491
RFR, ETR, XGBR	0.8204	0.8936	-150.2365	0.8428	0.7506	0.9028	0.6078	0.8787	0.8497	0.9027	0.8948	0.8937	0.85
RFR, ETR, BAR	0.8255	0.8897	-46.5583	0.8583	0.8129	0.9012	0.3715	0.8982	0.8556	0.9018	0.9068	0.9014	0.8606
RFR, GBR, XGBR	0.8212	0.8954	-69.2848	0.8362	0.7409	0.9028	0.8262	0.8768	0.8457	0.9015	0.9059	0.8981	0.8508
RFR, GBR, BAR	0.827	0.8869	-18.8438	0.8604	0.7929	0.9013	0.828	0.8935	0.8579	0.903	0.905	0.8991	0.8621
RFR, XGBR, BAR	0.827	0.8877	-43.126	0.8564	0.7805	0.9035	0.3474	0.8949	0.8639	0.9037	0.9051	0.9001	0.8636
ETR, GBR, XGBR	0.8266	0.896	-372.123	0.8454	0.7276	0.9038	0.8272	0.8857	0.846	0.9027	0.9046	0.8982	0.8473
ETR, GBR, BAR	0.8266	0.8833	-91.9138	0.8409	0.7826	0.9012	0.7691	**0.9**	0.859	0.9026	0.9113	0.902	0.8552
ETR, XGBR, BAR	0.821	0.8847	-138.3163	0.8726	0.7696	0.9038	0.2966	0.8927	0.8603	0.9038	0.9083	0.9015	0.8593
GBR, XGBR, BAR	0.827	0.8843	-76.3932	0.8563	0.7585	0.9039	-3.8396	0.8972	0.86	0.9029	0.8971	0.9008	0.8587
RFR, ETR, GBR, XGBR	0.821	0.8936	-847.4178	**0.8301**	0.7462	0.9032	-1.024	0.8783	0.8498	0.9018	0.9014	0.894	0.8518
RFR, ETR, GBR, BAR	0.8255	0.8871	-1858.1811	0.8605	0.7931	0.9015	0.7607	0.8973	0.8579	0.903	0.8923	0.9019	0.8612
RFR, ETR, XGBR, BAR	0.8255	0.8878	-2247.591	0.8587	0.7836	0.9038	-1.2588	0.8983	**0.8639**	0.9037	0.9059	0.9013	0.8638
RFR, GBR, XGBR, BAR	0.8261	0.8875	-385.7334	0.8581	0.774	0.9041	-3.4404	0.8929	0.857	0.903	0.8892	0.9012	0.8625
ETR, GBR, XGBR, BAR	0.826	0.8842	-5965.465	0.8659	0.7636	0.9046	-1.7782	0.8996	0.8607	0.903	0.8857	0.902	0.8557
RFR, ETR, GBR, XGBR BAR	0.8251	0.8876	-10960.3019	0.8576	0.7758	**0.9046**	-20.4833	0.8969	0.856	0.903	0.8806	0.9019	0.8612

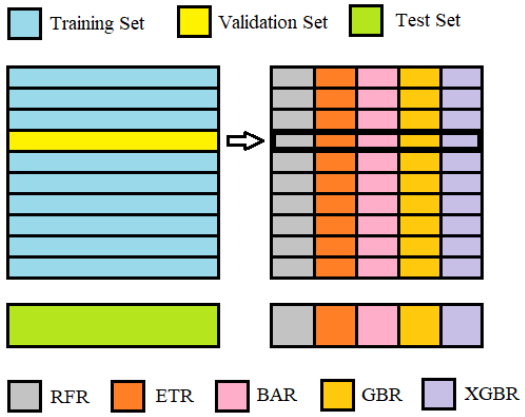

Fig. 7. Pictorial Representation of Stacking based prediction

4) Stacking (Repeated K-Fold Cross Validation): The process used is similar to Section IV-C3, except instead of using K-Fold Cross Validation, Repeated K-Fold Cross Validation has been used where the number of repeats is 10. Hence the newly created training set will be ten times larger than the training set created in Section IV-C3.

D. Performance Analysis of the developed Hybrid Ensemble Regression Models

The performance analysis of the developed Hybrid Ensemble Regression Models is given Tables IV and V. Table IV shows MSE values of the developed models, and Table V shows the EVS values. The best value among the different combinations for each of the developed models are highlighted. The number of models used from Section IV-B is five (RFR, ETR, BAR, GBR and XGBR), hence the number of combinations available for analysis is 31.

It can be observed from Table IV and Table V that among the models used to predict from the newly formed train and test set, PR[4] performed best in blending, stacking (repeat = 1) and stacking (repeat = 10) followed by MLR, DTR[6] and SVR[L]. The MSE values of SVR[L] in all the techniques are ten times higher than the other methods. No matter what algorithm is used to predict the newly formed train set and test set stacking (repeat = 10) performed best followed by stacking (repeat = 1) and blending. For blending with PR[4], most of the combinations are have negative EVS values indicating that the models are unstable as they are overfitting.

TABLE VI
OBSERVATIONS FROM HYBRID ENSEMBLE REGRESSION ANALYSIS

Methods	Performance Measures	Values	Combinations
Simple Averaging	MSE	0.000513	GBR, BAR
	EVS	0.8285	GBR, BAR
Blending [PR[4]]	MSE	0.000298	XGBR
	EVS	0.9036	XGBR
Stacking (Repeat = 1) [PR[4]]	MSE	0.000291	XGBR
	EVS	0.9059	XGBR
Stacking (Repeat = 10) [PR[4]]	MSE	**0.000274**	**ETR, BAR**
	EVS	**0.9113**	**ETR, BAR**

Major observations in Table IV and Table V are tabulated, and are shown in Table VI. It can be clearly observed from Table VI that, for all the four hybrid ensemble regression models the same combination of ensemble regression models gives the minimum MSE and the maximum EVS values, and the best result is produced by Stacking (Repeats = 10) having PR[4] for combining the results of ETR and BAR.

E. Comparison between the models based on Actual Rainfall values

The models developed in this paper and [3] showed promising results. However, the results were based on normalised values of rainfall. Hence a comparison has been made between the models by converting the normalised rainfall values to its original state, i.e. the maximum and minimum values of rainfall before the normalisation was taken and the normalised predicted result was converted based on (5).

$$A_i = A'_i * (A_{max} - A_{min}) + A_{min} \tag{5}$$

The explanation for the symbols in (5) is given in Section III-B. Since the maximum and minimum values of the predicted values of rainfall is unknown, the values belonging to the actual rainfall are used. The graphical representation of the comparison between the models is given in Fig. 6.

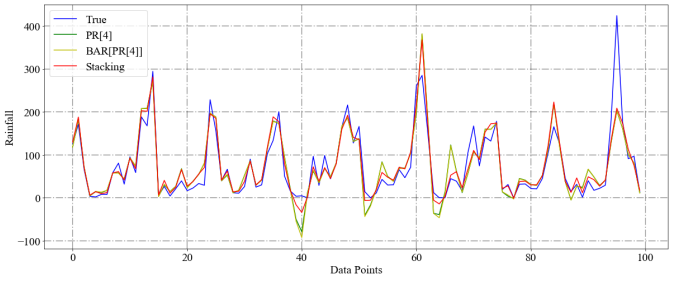

Fig. 8. Comparison between the models based on actual rainfall values

It can be observed from Fig 6 that all the developed models predict rainfall with less error and capture the variation. It can also be observed that the line of stacking is very close with the actual rainfall line followed by the other two models. The gap between the PR[4] and BAR[PR[4]] is very less. The comparison between the best model in this paper and the best models from other papers using the same dataset is shown in Table VII.

TABLE VII
COMPARISON OF PROPOSED MODEL RESULT WITH OTHER PAPERS USING SAME DATASET

Paper Name	MSE	RMSE	EVS
S. K. Mohapatra et al. in [1]	-	9.2433	0.8473
A. H. Manek et al. in [2]	-	0.2060	-
P. Ganesh et al. in [3]	0.00052	0.0227	0.8267
Proposed Model	**0.000274**	-	**0.9113**

It can be observed from Table VII that the proposed model has better EVS values than the models in [1], [3]. The RMSE

value of the model in [3] is better than the models in [1], [2] and the MSE value of the proposed model is better than the model in [3]. Hence it can be concluded that the proposed model performs better than the other models in Table VII.

V. Conclusion

In this paper, we have developed various ensemble regression models using methods such as Bagging and Boosting to predict rainfall in districts belonging to the state of Tamil Nadu, India. Based on preliminary analysis, it was concluded that Bagging Regression with the number of estimators as 70 and base estimator as Polynomial Regression with the degree as four performs best. Simple averaging, Blending and Stacking were used to predict predicted results of ensemble regression models, and the developed model was called as the hybrid ensemble regression model. On analysis it was concluded that using ensemble technique Stacking with repeated k-fold cross validation where the number of folds is 10 and the number of repeats is 10, having the base model as Extra Trees Regression (with the number of estimators as 90 and maximum depth as 11) and Bagging Regression (with the number of estimators as 90 and maximum depth as 11). It was concluded that the model mentioned above performed two times better than the Hybrid Ensemble Regression Model using Simple Averaging and all the Ensemble Regression Models.

References

[1] Sandeep Kumar Mohapatra, Anamika Upadhyay, and Channabasava Gola. Rainfall prediction based on 100 years of meteorological data. In *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, pages 162–166. IEEE, Oct 2017.

[2] Aishwarya Himanshu Manek and Parikshit Kishor Singh. Comparative study of neural network architectures for rainfall prediction. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pages 171–174. IEEE, Jul 2016.

[3] Preetham Ganesh, Harsha Vardhini Vasu, and Dayanand Vinod. Forecast of Rainfall Quantity and its Variation using Environmental Features. In *2019 Innovations in Power and Advanced Technologies (i-PACT)*. IEEE, 2019, in press.

[4] Cristian Valencia-Payan and Juan Carlos Corrales. A multiscale based rainfall amount prediction using multiple classifier system. In Plamen Angelov, Jose Antonio Iglesias, and Juan Carlos Corrales, editors, *Advances in Information and Communication Technologies for Adapting Agriculture to Climate Change*, pages 16–28, Cham, 2018. Springer International Publishing.

[5] Amit Kumar Sharma, Sandeep Chaurasia, and Devesh Kumar Srivastava. Supervised rainfall learning model using machine learning algorithms. In Aboul Ella Hassanien, Mohamed F. Tolba, Mohamed Elhoseny, and Mohamed Mostafa, editors, *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, pages 275–283, Cham, 2018. Springer International Publishing.

[6] V.P Tharun, Ramya Prakash, and S. Renuga Devi. Prediction of Rainfall Using Data Mining Techniques. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 1507–1512. IEEE, Apr 2018.

[7] Andrew Kusiak, Xiupeng Wei, Anoop Prakash Verma, and Evan Roz. Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):2337–2342, Apr 2013.

[8] Kesheng Lu and Lingzhi Wang. A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction. In *2011 Fourth International Joint Conference on Computational Sciences and Optimization*, pages 1343–1346. IEEE, Apr 2011.

[9] Rashmi Priya and Dharavath Ramesh. Adaboost.rt based soil n-p-k prediction model for soil and crop specific data: A predictive modelling approach. In Anirban Mondal, Himanshu Gupta, Jaideep Srivastava, P. Krishna Reddy, and D.V.L.N. Somayajulu, editors, *Big Data Analytics*, pages 322–331, Cham, 2018. Springer International Publishing.

[10] Junliang Fan, Xiukang Wang, Lifeng Wu, Hanmi Zhou, Fucang Zhang, Xiang Yu, Xianghui Lu, and Youzhen Xiang. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in china. *Energy Conversion and Management*, 164:102 – 111, 2018.

[11] Haoyuan Hong, Junzhi Liu, Dieu Tien Bui, Biswajeet Pradhan, Tri Dev Acharya, Binh Thai Pham, A-Xing Zhu, Wei Chen, and Baharin Bin Ahmad. Landslide susceptibility mapping using j48 decision tree with adaboost, bagging and rotation forest ensembles in the guangchang area (china). *CATENA*, 163:399 – 413, 2018.

[12] Muhammad Waseem Ahmad, Jonathan Reynolds, and Yacine Rezgui. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production*, 203:810 – 821, 2018.

[13] Alberto Torres-Barran, lvaro Alonso, and Jose R. Dorronsoro. Regression tree ensembles for wind energy and solar radiation prediction. *Neurocomputing*, 326-327:151 – 160, 2019.

[14] Harikrishnan N B, Vinayakumar R, and Soman K P. A machine learning approach towards phishing email detection cen-security@iwspa 2018. Mar 2018.

[15] K. Lakshmi Devi, P. Subathra, and P. N. Kumar. Tweet sentiment classification using an ensemble of machine learning supervised classifiers employing statistical feature selection methods. In Vadlamani Ravi, Bijaya Ketan Panigrahi, Swagatam Das, and Ponnuthurai Nagaratnam Suganthan, editors, *Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO - 2015)*, pages 1–13, Cham, 2015. Springer International Publishing.

[16] H. Liang, L. Song, and X. Li. The rotate stress of steam turbine prediction method based on stacking ensemble learning. In *2019 IEEE 19th International Symposium on High Assurance Systems Engineering (HASE)*, pages 146–149, Jan 2019.

[17] P. Disorntetiwat and C. H. Dagli. Simple ensemble-averaging model based on generalized regression neural network in financial forecasting problems. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pages 477–480, Oct 2000.

[18] Iwan Syarif, Ed Zaluska, Adam Prugel-Bennett, and Gary Wills. Application of bagging, boosting and stacking to intrusion detection. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 593–602, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[19] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996.

[20] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, Aug 1995.

[21] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, Apr 2006.

[22] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[23] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.