Title	Link	Year	# of citations	Assigned To	About the dataset (data source?, daily/weekly/monthly?, how many samples?)	Methodology (how was the data pre-processed?, what graphs/statistics did they use to represent the data?, what was being predicted?, which ML algos applied?, anything else that is relevant)	Results (rmse, accuracy, f1-score, confusion matrix etc. Anything else reported by the paper)	Comments
Weather Forecasting for the North-Western region of Bangladesh: A Machine Learning Approach	https://ieeexplore.ieee.org/abstract/document/9225389	2020	0	jamil	The dataset was collected from Bangladesh Metoroligical Depertment (BMD). It consisted of thirty years of daily data on weather. The data originated at the seven weather station situated at northwest part of bangladesh.	There weren't any significant work done on preprocessing the dataset, except for splitting the dataset into test and train set (test: 80%, train: 20%). Two ML algorithms ANN (Artifical Nural Networks), and ELM (Extream Learning) were used for forcasing rain, humidity, wind pressure and temperature. ELM performed better than ANN.	The performence of the algorithms were measured using three error metrics, a performence parameter and correlation cofficient. They were: MAE (mean absolute error), RSME (root mean square error), MASE (mean absolute scaled error), PP (performence parameter) and CC(correlation cofficient). Prediction performence for rainfall on Dataset (2007-2017). MAE(%): ANN-6.189; ELM-3.075 RMSE(%): ANN-0.212; ELM-0.100 PP(%): ANN-0.678; ELM-0.715 CC: ANN-.979; ELM-.551 Prediction performence for Temperature on Dataset (2007-2017). MAE (%): ANN-0.512; ELM-0.900 RMSE (%): ANN-0.52; ELM-0.100 PP: ANN-0.678; ELM-0.715 CC: ANN-0.979; ELM-0.551 Prediction performence for Wind Pressure on Dataset (2007-2017) MAE (%): ANN-1.149; ELM-3.075 RMSE (%): ANN-0.212; ELM-0.100 PP: ANN-0.678; ELM-0.715 CC: ANN-0.979; ELM-0.557 Prediction performence for Humidity on Dataset (2007-2017) MAE (%): ANN-2.156; ELM-2.045 RMSE (%): ANN-0.501; ELM-1.643 PP: ANN-0.287; ELM-0.767 CC: ANN-0.279; ELM-0.857	

Title	Link	Year	# of citations	Assigned To	About the dataset (data source?, daily/weekly/monthly?, how many samples?)	Methodology (how was the data pre-processed?, what graphs/statistics did they use to represent the data?, what was being predicted?, which ML algos applied?, anything else that is relevant)	Results (rmse, accuracy, f1-score, confusion matrix etc. Anything else reported by the paper)	Comments
Effectiveness of Ensemble Machine Learning Algorithms in Weather Forecasting of Bangladesh	https://link.springer.com/chapter/10.1007/978-3-030-73603-3_25	2021	1	rifat	The raw dataset was collected from BMD (Bangladesh Meteorological Division) for the year 2012 to 2018. The dataset consists of four primary attributes of daily: wind speed, humidity, temperature, and rainfall collected at 33 weather stations across Bangladesh. The data for the years 2012 to 2017 was used for training, while the data of 2018 was used for testing.	The collected raw dataset required preprocessing and a lot of cleaning to form a structured dataset. As in regression problems, the variables should be centered so that the predictors can achieve a mean of “0” value. This paper applied the standardization scaling technique to make the values centered around the mean with a unit standard deviation, making the attribute zero and the resultant distribution unit standard deviation.[No graphs were provided to show the data distribution] The paper focuses on predicting weather forecasts (high-low temperature, rainfall) monthly or daily. Several ensemble regression algorithms, including support vector regression (SVR), linear regression, Bayesian ridge, gradient boosting (GB), extreme gradient boosting (XGBoost), category boosting (CatBoost)*, adaptive boosting (AdaBoost), k-nearest neighbors (KNN), and decision tree regressor (DTR)* were applied to the dataset.	MAE, MSE, MAPE scores were considered to evaluate the performance of those models. DTR performed exceedingly well in forecasting rainfall with MAE: 5.17, MSE: 944.19 and MAPE: 2.73%.	
Prediction of Temperature and Rainfall in Bangladesh using Long Short Term Memory Recurrent Neural Networks	https://arxiv.org/abs/2010.11946	2020	1	ferdous	Monthly temparature and rainfall data of 115 years (1901-2015) of Bangladesh from kaggle.	Predicted rainfall level and temparature. 2014-2015 data taken for testing and rest for training. LSTM was used to predict rainfall and temparature.	Mean error for rainfall prediction=-17.64mm	
An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives	https://www.sciencedirect.com/science/article/abs/pii/S0957417417303457	2017	75	ferdous	Daily rainfall data of 42 cities across Europe and USA from year 1990 to 2010. The writing suggests that no other features were used apart from rainfall.	Two version of datasets were created (i) one with raw daily rainfall and (ii) another with 'sliding window accumulated' (time-series smoothing technique) rainfall value. Citywise daily rainfall graphs were used to show how chaotic the daily rainfall data was, and analysing seasonal wet and dry periods. 6 ML algorithms: Genetic Programming, Support Vector Regression, Radial Basis Neural Networks, M5 Rules, M5 Model trees, and k-Nearest Neighbours were compared with traditional 'Markov chain extended with rainfall prediction' (MCRP).	Citywise RMSE was reported for each of the ML algos. SVR (kernel=RBF) had best 'mean score' RMSE with 1.90 for US and 2.50 for Europe. Showed a graph with predicted rainfall overlayed on top of actual rainfall data to present performance of the models.	

Title	Link	Year	# of citations	Assigned To	About the dataset (data source?, daily/weekly/monthly?, how many samples?)	Methodology (how was the data pre-processed?, what graphs/statistics did they use to represent the data?, what was being predicted?, which ML algos applied?, anything else that is relevant)	Results (rmse, accuracy, f1-score, confusion matrix etc. Anything else reported by the paper)	Comments
Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods	https://www.sciencedirect.com/science/article/abs/pii/S0022169420302493	2020	11	rifat	The rainfall data were collected from CIATF (maintains a database of rainfall observations). Their database contained 125 gauges, combining daily and sub-daily values, with an average coverage of 15 years. The gauges with the best information were considered in the analysis. All stations with at least 25 years of data from 1979-2015 were selected for this research.	[No relevant information about how the data were preprocced] A table showing the list of gauges and their average attribute values- used in the study was provided. Six different machine learning methods: LR, RF, KNN, SVM, K-Means, NN* were used in this paper to predict rainfall. A table depicting the model's hyperparameters and their values which were tested in the model's optimization step using the 'Grid Search,' was provided. During the CV procedure, several sets of hyperparameters were tested to find the optimal one. The performance of all the tested hyperparameters sets, both for classifiers (f-score) and regressors (R score), were shown in a figure.	In this paper, f-score and R were considered to evaluate the performance skill for the whole combinations of hyperparameters. RMSE and R were considered to evaluate overall model performances. NN, with an average f-score close to 0.4 and average R score of 0.37, appears as the best performing model (closely followed by the f scores of LR: 0.37 and SVM: 0.36).	
Rainfall Prediction using Machine Learning & Deep Learning Techniques	https://ieeexplore.ieee.org/abstract/document/9155896	2020	6	ferdous				waiting on response to full-text request on researchgate
Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia	https://www.sciencedirect.com/science/article/pii/S2090447920302069	2021	10	~~jamil~~ ferdous	Data collected from 10 stations in Malaysia from 2010 to 2019 with a total of 3455 daily rainfall instances.	Two version of daily, weekly, monthly and 10-days rainfall datasets were created. One using Autocorrelation Function (ACF) and the other using Projected Error as features (?). The Projected Error datasets were normalized using: MinMax, LogNormal and ZScore separately. The forecasting model uses four different ML algorithms, which are Bayesian Linear Regression (BLR), Boosted Decision Tree Regression (BDTR), Decision Forest Regression (DFR) and Neural Network Regression (NNR).	In ACF method, BDTR model gives the best R2-score. For daily rainfall 0.9739693, for weekly rainfall 0.989461, for 10 days rainfall 0.9894429 and for monthly rainfall 0.9998085. In Projected Error method, best model BDTR for daily R score 0.737978 and for weekly rainfall error prediction R-2 score 0.7921. For monthly rainfall error prediction DFR outperformed other models with R-2 score 0.7623. Ffor 10-days rainfall error prediction, NNR model with ZScore normalization outperformed other models with R-2 score 0.61728	
Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches	https://www.nature.com/articles/s41598-020-67228-7	2020	44	ferdous	115 years (1901-2015) annual rainfall data of 34 regions in India.	Mann-Kendall test was used to detect long term rainfall trends, Pettitt & Standard Normal Homogeneity test was used to detect seasonal and annual abrupt changes in rainfall. Multi layered perceptron (ANN) was used to predict rainfall of future years from 2015-2030 using data of 1901-2015.	RMSE, MAE was used for evaluation but exact values was not reported -__-	

Title	Link	Year	# of citations	Assigned To	About the dataset (data source?, daily/weekly/monthly?, how many samples?)	Methodology (how was the data pre-processed?, what graphs/statistics did they use to represent the data?, what was being predicted?, which ML algos applied?, anything else that is relevant)	Results (rmse, accuracy, f1-score, confusion matrix etc. Anything else reported by the paper)	Comments
A Novel Machine Learning Based Approach for Rainfall Prediction	https://link.springer.com/chapter/10.1007/978-3-319-63673-3_38	2017	2	rifat	In this paper, the authors did not mention the source of the dataset. However, the dataset consisted of the following attributes: temperature, cloud fraction, wind speed, humidity, rainfall.	For data processing, they removed the missing values and replaced them with center measured values like mean. After that, they normalized the values between [0...1] to minimize the overall error. In this paper, they used hybrid intelligent system data mining consists of the combination of ANN with GA-MLP (Graph-Augmented Multi-Layer Perceptrons) for optimizing neural network parameters to predict rainfall.	They mentioned considering the MSE and RMSE scores to evaluate the performance but did not provide any result/evaluation data table. They trained the model until they accomplished a minimum error of 0.05%.	

Title	Link	Year	# of citations	Assigned To	About the dataset (data source?, daily/weekly/monthly?, how many samples?)	Methodology (how was the data pre-processed?, what graphs/statistics did they use to represent the data?, what was being predicted?, which ML algos applied?, anything else that is relevant)	Results (rmse, accuracy, f1-score, confusion matrix etc. Anything else reported by the paper)	Comments
Predicting Rainfall using Machine Learning Techniques	https://arxiv.org/abs/1910.13827	2019		jamil	The data was collected from several weather stations in Australia. It is also available in kaggle in following link (https://www.kaggle.com/jsphyg/weather-dataset-rattle-package). The dataset consisted of daily weather observation.	The data was first speculated using Univariate Visualization and Correlation Heat Map. During the specalutaion the percentage of null values in the features and the class imbalance issue was discovered. The null values was replaced with the mean values of respective features. The class imbalance issue was tackled using oversampling and undersapling techniques. Feature selection was done by Univariate Selection using sklearn library's **SelectKBest** class. Logistic Regression, Decision Tree, KNN, Random Forest, AdaBoost and Gradient Boosting were the ML algorithms that were used.	Accuracy, Area Under Curve (auc), Pricision, Recall, F1 Score and Confusion Matrix were used to evaluate the ML algorithms. ***Over Sampled Data:*** **Accuracy** Gradient Boosting- 0.25: ~85% Gradient Boosting-0.05: ~85% Gradient Boosting-0.1: ~85% Logistic Regression: ~84% Ensemble adaBoost: ~83% Random Forest: ~83% KNN-29: ~83% KNN-27: ~83% KNN-25: ~83% Decision Tree: ~81% **AUC** Gradient Boosting- 0.25: ~86.1% Gradient Boosting-0.05: ~85.8% Gradient Boosting-0.1: ~86.1% Logistic Regression: ~84.1% Ensemble adaBoost: ~85.8% Random Forest: ~88.2% KNN-29: ~87.3% KNN-27: ~87.4% KNN-25: ~87.2% Decision Tree: ~70.9% ***Under Sampled Data:*** **Accuracy** Gradient Boosting- 0.25: ~78% Gradient Boosting-0.05: ~77% Gradient Boosting-0.1: ~78% Logistic Regression: ~77% Ensemble adaBoost: ~77% Random Forest: ~76% KNN-29: ~76% KNN-27: ~75.5% KNN-25: ~75.5% Decision Tree: ~71% **AUC** Gradient Boosting- 0.25: ~82.3% Gradient Boosting-0.05: ~82.8% Gradient Boosting-0.1: ~82.5% Logistic Regression: ~84.2% Ensemble adaBoost: ~82.4% Random Forest: ~82.4% KNN-29: ~82.1% KNN-27: ~81.8% KNN-25: ~81.8% Decision Tree: ~72.8%	

Title	Link	Year	# of citations	Assigned To	About the dataset (data source?, daily/weekly/monthly?, how many samples?)	Methodology (how was the data pre-processed?, what graphs/statistics did they use to represent the data?, what was being predicted?, which ML algos applied?, anything else that is relevant)	Results (rmse, accuracy, f1-score, confusion matrix etc. Anything else reported by the paper)	Comments
Application of machine learning to an early warning system for very short-term heavy rainfall	https://www.sciencedirect.com/science/article/abs/pii/S0022169418309211	2019	23	ferdous	Hourly meteorological data from 652 automatic weather stations in South Korea from 2007 to 2012. The dataset had 23 features. The average number of instances for each station is 46200. The dataset was HIGHLY IMBALANCED, with 99% instances belonging to one class (binary classification).	A binary classification problem to predict heavy rain would occur 3h prior. Experimented with pre-processing by (i) 'Selective' discretization and (ii) Principal Component Analysis (PCA). Logistic Regression, ANN, 1-nearest neighbor, C4.5, random forests, support vector machines, SMO (sequential minimal optimization), and RIPPER (repeated incremental pruning to produce error reduction) was used with 3 fold cross validation.	F-measure and Equitable Threat Score (ETS) was reported (because data was HIGHLY IMBALANCED). Logistic Regrssion with both pre-processing performed best with f1-score 0.46, and ETS of 0.30.	find out more about the feature *extraction* technique: Principal Component Analysis (PCA)
A Data-Driven Approach for Accurate Rainfall Prediction	https://ieeexplore.ieee.org/abstract/document/8789447	2019	6	ferdous	Precipitable Water Vapor (PWV) and other weather data collected from Nanyang Technological University Station (2012-2015) and Singapore National University Station (2016). The weather parameters are recorded every 1 minute. Rain : No rain = 1 : 104	Binary classification of rain or no rain was done with 5 minute lead time taking into account 30 minutes of data. Sine and cosine of hour of day, day of month were taken as seasonal ,diurnal features. Down sampling was applied to address the massive imbalance. Only SVM was used. Experimentation was conducted by including and excluding each one of the features.	TPR=80.4%, FPR=20.3%, and Accuracy=79.6%.	
All India summer monsoon rainfall prediction using an artificial neural network	https://link.springer.com/article/10.1007/s003820050328	2000	181	rifat	The time series dataset contains the summer monsoon rainfall data for the period 1871-1994. The preprocessed dataset was obtained from another research project.	The purpose of their paper was to predict the rainfall of t+1 year, using the rainfall data of t, t-1, t-2, t-3, t-4 years. ANN, containing four layers with two hidden layers, was used to predict the summer monsoon rainfall.	For making a robust comparison of the observed and predicted values of rainfall, they considered various statistics: Mean (pred): 829.5 (observed: 840) SD (pred): 69.2 (observed: 90.4) CC (corr coeff): 0.81 RMSE: 54.24 PP (performance parameter): 0.36 PC (percent correct): 50 HSS (Heidke skill score): 0.31	
Rainfall Prediction: A Deep Learning Approach	https://link.springer.com/chapter/10.1007/978-3-319-32034-2_13	2016	27	jamil	The dataset was collected from meteorological stations in Manizales, Colombia. It consisted of daily data from year 2002 to 2013. There were 4216 instances isntances in the dataset.	A deep learning architecture was used to predict the accumulated rainfall for the next day. The architecture consisted of an autoencoder network at the input and a multilayered perceptron at the output. The autoencoder was made up of three layers: the input layer, the hidden layer(sigmoid activation function, and the output layer). It was used to extract the non-linear features of the data	Mean Squared Error(MSE) and Root Mean Squared Error was used to evaluate the architecture. MSE Autoencoder and MLP:40.11 MLP: 42.34	