Effectiveness of Ensemble Machine Learning Algorithms in Weather Forecasting of Bangladesh

Atik Mahabub[1], Al-Zadid Sultan Bin Habib[1,2], M. Rubaiyat Hossain Mondal[3],
Subrato Bharati[3], and Prajoy Podder[3](✉)

[1] Khulna University of Engineering and Technology, Khulna 9203, Bangladesh
[2] Jahangirnagar University, Savar, Dhaka 1342, Bangladesh
[3] Institute of ICT, Bangladesh University of Engineering and Technology,
Dhaka 1205, Bangladesh

Abstract. Machine learning (ML) is considered as a fundamental approach in predicting different phenomena including weather. This paper focuses on the application of ML models in weather forecasting of Bangladesh where the weather changes frequently. The novelty of this work is in the application of ensemble regression algorithms to a raw dataset collected from Bangladesh Meteorological Division for the year 2012 to 2018. The dataset has different attributes, including wind speed, humidity, temperature, and rainfall collected at 33 weather stations across Bangladesh. The dataset is split into training and testing portions; the data for the years 2012 to 2017 is used for training, while the data for the year 2018 is used for testing. The prediction is done using several ML-based 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). Our results show that the DTR and CatBoost algorithms outperform the algorithms reported in the literature in terms of mean squared error (MSE), mean absolute percentage error (MAPE), and mean absolute error (MAE).

Keywords: Rain forecasting · Environment · Bangladesh · Machine learning · Data mining

1 Introduction

Weather forecasting is supposed to be a prime factor for Bangladesh's economy as agriculture plays a vital role in the country's overall Gross Domestic Product (GDP) which accounts for approximately 20% of the total amount. Nearly 70% of its total population lives in rural areas and 60% of them earn their livelihood from village stuff. So, the discrepancy of rainfall, humidity, wind speed, the temperature in time, space, and aggregation affects the country's agriculture which might hamper the economy to a greater extent. Initially, meteorologists collect quantitative data to assemble the forecast.

A. Abraham et al. (Eds.): IBICA 2020, AISC 1372, pp. 267–277, 2021.
https://doi.org/10.1007/978-3-030-73603-3_25

Apart from its significant importance in agriculture and economy, a successful forecast of rainfall and thunderstorm can save airlines from falling into accidents or in case of unwanted flight delays or unwanted crashes which may cause the death of many people [1–5].

A few types of data sources like satellite, marine, land-based stations, weather balloons, radar, and paleoclimatic are available and different sorts of instruments are used to collect the data for measurement purposes. After completing the final measurement, the data is sent to the satellite from ground weather stations [6]. Intelligent weather prediction techniques can help us to a certain degree that can help us to make effective decisions that can save valuable lives, times, and property at a time. With the passage of time, science and technology have advanced to the next level, and weather pattern discovery has attracted more attention. It involves the anticipation of how the current circumstance with the air will change in which current climate conditions are taken via ground discernments, e.g. boats, radar, satellite, aeroplanes, etc. Then the accumulated data is forwarded to the meteorological department for further analysis and processing, which results in knowledge representation via charts, graphs, or even maps. Algorithms trade a large number of discernments onto the surface and upper-air maps and draw the lines on the maps with cooperation from meteorologists. Later the approximate look of the map will be determined by the algorithms. These sorts of weather forecasting using algorithms are delineated as numerical or computational climate forecasting [7–9]. Despite having several weather forecasting techniques, complex physics behind weather does not make it a simpler task which depends on countless traits, and which is also a turbulent and perplexing climatic event. Intelligent devices can be helpful to collect data, and for further analysis, cognitive tools are always used [10, 22, 23]. Moreover, human-made events also play a vital role to affect the parameters of weather. Multiple cognitive methods, including artificial neural network (ANN), genetic algorithm (GA) were applied to predict the weather updates [11, 12].

A number of research work has applied machine learning (ML) based regression models to predict the weather. However, only a few works have reported the successful application of ensemble ML algorithms on a Bangladeshi weather dataset. This work focuses on this issue. The main contributions of this work are as follows.

1) A Bangladeshi weather dataset having different attributes including rainfall and temperature for the year 2012 to 2018 is split into training and testing portions; the data for the years 2012 to 2017 is used for training, while the data for the year 2018 is used for testing.
2) 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) are applied to the Bangladeshi dataset. Compared to the existing methods reported in the literature, DTR and CatBoost algorithms show better computational performance both in training and testing steps.

The rest of the paper is prepared as follows: Sect. 2 presents the overall methodology of this research work which includes data preprocessing and splitting of the data samples into training and testing portions. Section 3 evaluates the performance of our ensemble algorithms in predicting rainfall, low temperature, and high temperature. Finally, Sect. 4 provides concluding remarks.

2 Methodology

The collected raw dataset from BMD has been preprocessed, and a lot of cleaning is required to convert it from a semi-structured dataset to a structured dataset and prepared for the implementation of our desired model. We have considered the data of the period of 2012 to 2018 as we want to obtain the possible suitable output for Bangladesh for the most recent weather data. After completing the preprocessing of the dataset initially we have prepared two separate datasets in CSV file, one for 2012–2017, for training and testing the model and another one for 2018 to make the forecasting of rain where each dataset has four specific parameters including wind speed, rainfall, humidity, and temperature (low and high). We have conducted our research on this preprocessed dataset, and later we have compared all the outcomes for better predictions. Provided data has the category as rainfall (millimetre), humidity (percentage of water in the air), wind speed (kilometre per hour), temperature (degree Celsius). The weather data is collected from 33 weather stations across Bangladesh which stations are considered the core government-controlled weather station of Bangladesh. Each data file has a similar column structure having a year, month, day, and other corresponding parameter values. Long-range forecasting can be divided into four categories, (a) periodicity approach, (b) correlation approach (c) extended synoptic approach, and (d) dynamical approach [13]. BMD is a government agency for weather prediction in Bangladesh. In 2007 BMD first introduced a statistical forecast system based on the ensemble technique [14–17]. Although their predictions were acceptable, that was always dependent on some specific predictors. However, we have tried to put it one-step forward through our research. Some ML algorithms (i.e., KNN, XGBoost, GB, DTR, AdaBoost, SVR, linear regression, Bayesian ridge, and CatBoost) are delicate to feature scaling whether that utilizes gradient descent as an optimization approach requires the data to be mounted. In regression problems, the variables should be centred so that the predictors can achieve a mean of “0” value. In this research work, the standardization scaling technique has been applied where the values are centred around the mean with a unit standard deviation which makes the attribute zero and the resultant distribution has a unit standard deviation [18]. After preprocessing and dataset cleaning to make it a structural dataset, the raw dataset is splitted into two parts, one for training data and another one for testing data. The model will be trained up using these regression-based learning algorithms in the training dataset. Next, the performance will be tested for the testing dataset for specific algorithms to match its learning and prediction level compared to the training dataset. In the first case of training and testing the model, we have used the first dataset file, which contains the weather data for the years between 2012 and 2017. Moreover, we have used the second dataset file containing weather data of 2018 to forecast the rain and match it with the collected data. In the end, the final output will be monitored by the Performance Evaluation Unit.

It has been reported in the literature [19–21] that the performance of an algorithm varies depending on the datasets and on the features considered. Based on the performance of each algorithm, the best algorithm has been selected for each problem or each

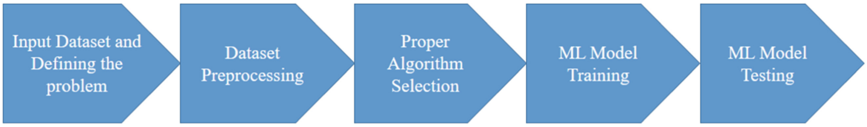

Input Dataset and Defining the problem
Dataset Preprocessing
Proper Algorithm Selection
ML Model Training
ML Model Testing

Fig. 1. The workflow of the ML Unit.

weather forecasting parameter. Figure 1 is the complete pictorial view of the ML Unit. At first, the dataset is entered, and the problem is defined. Theoretically, defining problems means what we need to do with the raw data; what we want to calculate, what kind of knowledge or features can be extracted. As weather forecasting from raw weather data is a regression type problem, the goal has been fixed to predict the weather. Here, at first, we have decided what to evaluate from these data.

Initially, four weather parameters have been selected to forecast rain using this model. Those four parameters are rainfall, wind speed, humidity, and temperature. Eventually, the temperature is predicted, splitting into two parts, e.g. high temperature and low temperature. Required preprocessing is done to prepare the dataset. The missing values are repaired, and the garbage values are removed via the preprocessing method. The proper algorithm is selected which is suitable for our data. The preprocessed data is split into two types, e.g. (i) training the dataset and (ii) testing dataset. Then the model is trained with the training dataset and finally, it is tested using the testing dataset. Usually, if it learns well from the training data then it provides better output for testing data too. The output performance is observed comparing the performance of the model with the testing dataset. If its performance does not deviate too much from the performance with training data, then its performance is considered satisfactory. Otherwise, the model or algorithm needs to be changed.

3 Performance Evaluations

Initially, the weather dataset is in CSV file, and Python 3.6 programming language has been used in Jupyter Notebook under Anaconda distribution for this work. After preprocessing, the dataset was split randomly into training (80%) and testing (20%) dataset to train the algorithms and to analyze their performances at a random split. To demonstrate the performance of the ML models, we set two distinct factors to measure: mean absolute error (MAE) and mean squared error (MSE) to check the performance of each algorithm. Apart from that, predictions were also represented graphically for each algorithm compared between the training data (represent as actual value) and testing data (represented by the predicted value).

Table 1 represents the statistical analysis of rainfall for the given dataset of the year 2012–2017. The values of MAE and MSE measure based on the testing dataset which indicates the learning efficiency of the ML models. From these measured values, we can notice that AdaBoost, linear regression and Bayesian ridge showed much more error compared to other algorithms in the case of predicting rainfall.

Table 1. Statistical analysis for rainfall for the 2012–2017 dataset

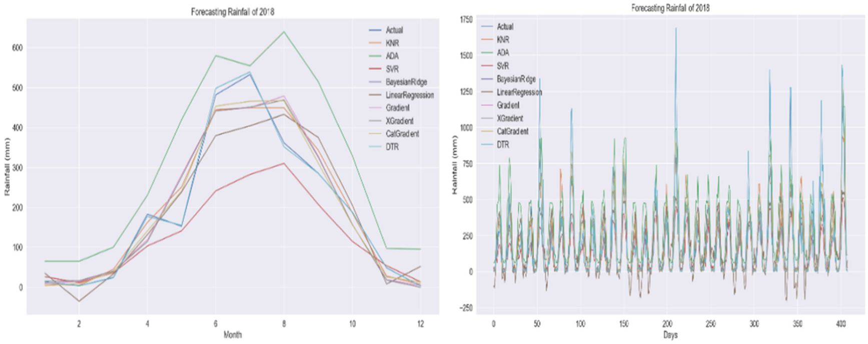
Algorithm	MAE	MSE	MAPE
KNN	81.41	17019.17	39.72%
XGBoost	73.20	13353.92	35.71%
GB	72.67	13208.35	35.46%
AdaBoost	132.29	28352.98	64.67%
DTR	95.78	33014.21	46.73%
SVR	71.88	15054.072	35.07%
Linear regression	124.30	29875.61	60.65%
Bayesian ridge	124.27	29873.45	60.63%
CatBoost	67.49	11657.74	32.93%

Table 2 is the statistical analysis of the rainfall in the case of forecasting for 2018. From the values of MAE, mean absolute percentage error (MAPE), and MSE, it can be noticed that DTR performs exceedingly well in forecasting. So, it is proved again that DTR plays a vital role in forecasting due to its structure and algorithm's tree construction.

Figure 2(a) and Fig. 2(b) illustrate the rainfall forecasting performance of our used ML algorithms. Here, Fig. 2(a) shows the monthly forecasting of rainfall for 2018 based on their previous experience from training and testing data. It clearly shows that AdaBoost does not show coherent performance along with other algorithms where SVR also lagged behind others in case of predicting the exact value. In the case of Fig. 2(b), we can notice the daily basis annual rainfall forecasting by our chosen ML models. It illustrates the combined performance of each algorithm in case of forecasting rainfall daily of the year. Both figures demonstrate the forecasting performance of our ML-based models, where the efficiency of the ML algorithms is visibly noticed. The output shows quite similar performance in the case of predicting, which proves the stability of the ML models.

Table 2. Statistical analysis for rainfall for the 2018 dataset

Algorithms	MAE	MSE	MAPE
KNN	87.33	23173.56	46.13%
XGBoost	81.51	19049.68	43.07%
GB	82.67	19480.34	43.67%
DTR	5.17	944.19	2.73%
AdaBoost	159.99	44107.75	84.54%
SVR	86.91	34323.64	45.91%
Linear regression	121.73	36339.32	64.31%
Bayesian ridge	121.67	36335.81	64.28%
CatBoost	69.89	14575.92	36.93%



(a) (b)

Fig. 2. Forecasting of rainfall for (a) each month (b) each day of the year for all algorithms compared to original values

Table 3. Statistical analysis of the model for high temperature for the 2012–2017 dataset

Algorithms	MAE	MSE	MAPE
KNN	0.785	0.97	2.65%
XGBoost	0.0015	4.815e−06	0.00509%
GB	0.000155	5.162e−08	5.23e−4%
DTR	0.0	0.0	0%
AdaBoost	0.2377	0.108	0.80%
SVR	0.199	0.109	0.67%
Bayesian ridge	1.158985e−11	2.0333e−22	3.91e−11%
Linear regression	2.0847e−14	6.0716e−28	7.033e−14%
CatBoost	0.426	0.3025	1.44%

Tables 3, 4, 5 and 6 indicate that ML algorithms show optimum output in case of predicting high and low temperature. Here, Table 3 and Table 5 set the standard for an initial dataset for testing data based on training data for high and low temperature, while Table 4 and 6 show performance for the second dataset for forecasting high and low temperatures. In both cases, the values of MAE, MAPE, and MSE were very negligible, which proved the authenticity of this approach of using ML-based algorithms in high and low temperature predicting. In a nutshell, it can be stated that ML algorithms provide an almost errorless prediction.

The MAE, MAPE, and MSE values were also significantly less than conventional methods which promise an opportunistic approach for high and low-temperature forecasting. All the algorithms we used in our research showed sheer corrosiveness which bolstered the claim of a most authentic high and low-temperature forecasting technique. Similarly, we can obtain the monthly and daily high and low-temperature forecasting observations from our developed ML models and compare them with the actual values

Table 4. Statistical analysis of the model for high temperature for the 2018 dataset

Algorithms	MAE	MSE	MAPE
KNN	0.7797	1.069	2.54%
XGBoost	0.00162	5.508e−06	5.29e−3
GB	0.000155	5.382e−08	5.05e−4%
DTR	0.0	0.0	0.0
AdaBoost	0.267	0.1325	0.87%
SVR	0.0319	0.0015	0.10%
Bayesian ridge	1.18e−11	2.14e−22	3.85e−11%
Linear regression	1.95e−14	5.67e−28	6.36e−14%
CatBoost	0.399	0.271	1.30%

Table 5. Statistical analysis of the model for low temperature for the 2012–2017 dataset

Algorithm	MAE	MSE	MAPE
KNN	0.896	1.37	4.04%
XGBoost	0.00063	4.74e−07	2.85e−3%
GB	0.000196	7.816e−08	8.83e−4%
DTR	0.0	0.0	0%
AdaBoost	0.4369	0.199	1.97%
SVR	0.03348	0.00177	0.15%
Bayesian ridge	1.16e−11	2.0779e−22	5.23e−11%
Linear regression	2.409e−14	7.843e−28	1.08e−13%
CatBoost	0.415	0.317	1.87%

Table 6. Statistical analysis of the model for low temperature for the 2018 dataset

Algorithm	MAE	MSE	MAPE
KNN	0.80	1.198	3.56%
XGBoost	0.00063	4.57e−07	2.78e−3
GB	0.00016	5.35e−08	7.10e−4%
DTR	0.0	0.0	0%
AdaBoost	0.36	0.16	1.62%
SVR	0.0399	0.0023	0.18%
Bayesian ridge	1.1676e−11	2.104e−22	5.14e−11%
Linear regression	2.0007e−14	6.4117e−28	8.81e−14%
CatBoost	0.3755	0.255	1.65%

which are preserved in the dataset of 2018. It can be noticed from Figs. 3, 4, 5 and 6 that the outputs of our used algorithms do not differ that much from the actual value, which ensures the forecasting accuracy of our used ML models. Figure 3 represents the monthly high-temperature forecast of 2018, whether Fig. 4 illustrates the high-temperature forecast for each day of 2018. Figure 5 depicts the monthly low-temperature forecast of 2018, whether Fig. 6 illustrates the low-temperature forecast for each day of 2018. From Figures, it can be stated that DTR showed the best performance in proving error less high-temperature forecasting. Apart from DTR, linear regression and Bayesian ridge show better performance compared to others.

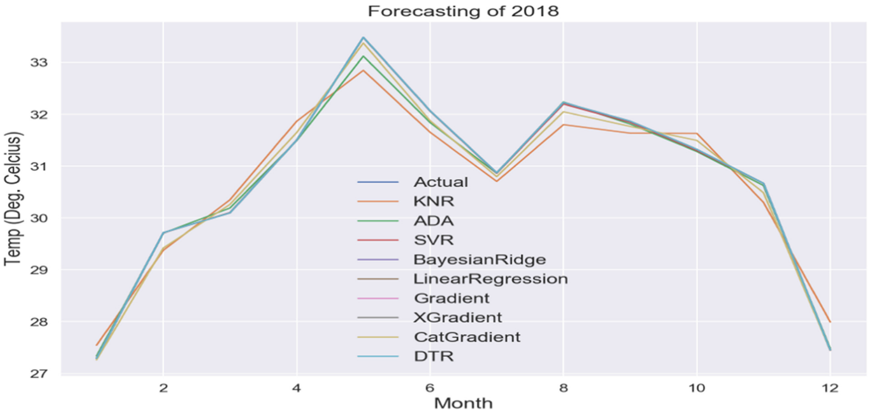


Fig. 3. Forecasting of high temperature for each month of the year for all algorithms compared to original values.

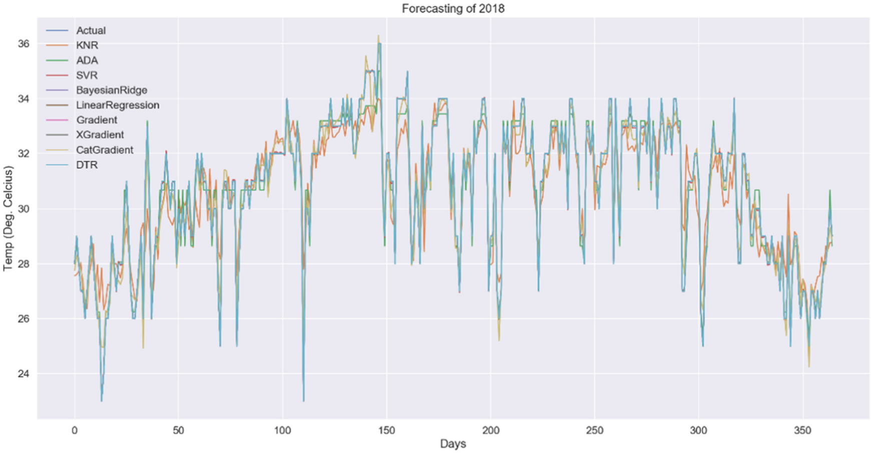


Fig. 4. Forecasting of high temperature for each day of the year for all algorithms compared to original values

The high and low-temperature prediction and forecasting here we also found that error is minimal compared to other conventional methods and DTR shows the best performance amongst all the algorithms. Once again, DTR provided almost zero error which established the claim of its being the best algorithm for weather forecasting parameters. Apart from that linear regression, Bayesian ridge and GB showed the least amount of error.

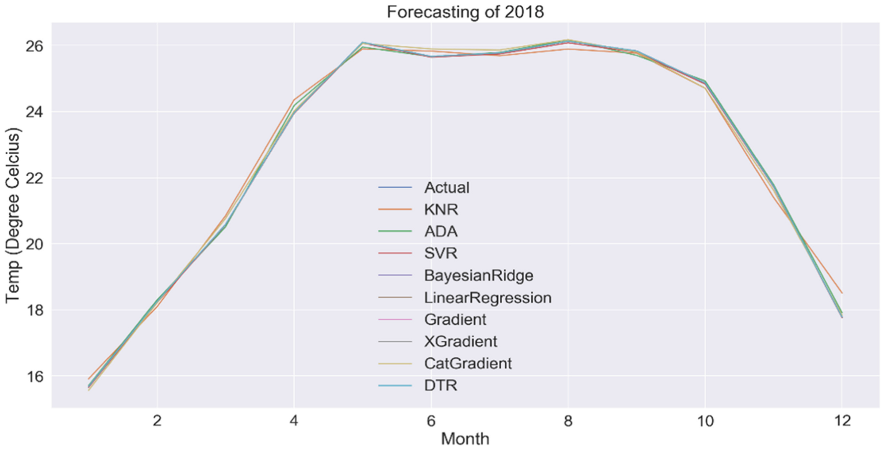


Fig. 5. Forecasting of low temperature for each month of the year for all algorithms compared to original values

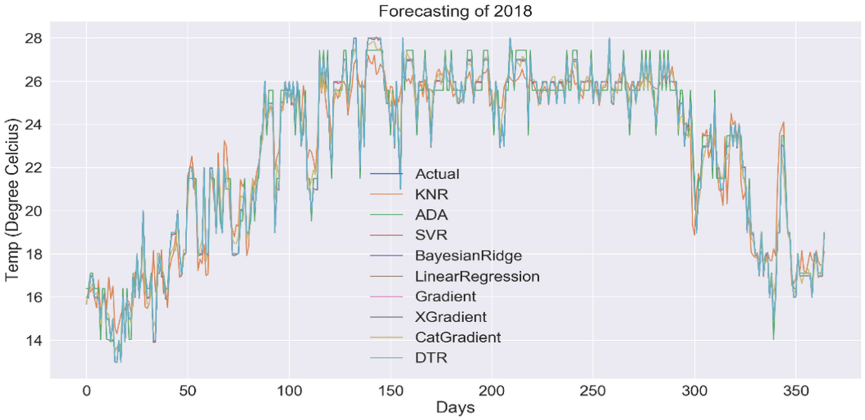


Fig. 6. Forecasting of low temperature for each day of the year for all algorithms compared to original values

4 Conclusions

In this paper, we have applied ensemble regression algorithms to forecast the rainfall, low temperature, and high temperature of Bangladesh. Our findings suggest that regression-based ML-algorithms are capable of predicting weather events with a narrow margin of error rate utilizing the smaller number of weather parameters. A number of algorithms are applied, and it is found that DTR and CatBoost methods have the best performance. However, the adaptability of DTR as a widespread nonlinear assumption makes it more ideal than CatBoost. The performance of a model depends on the dataset considered. Therefore, a number of other Bangladeshi datasets should be taken into consideration to validate the effectiveness of DTR and CatBoost algorithms. In future, weather forecasting can be done for other features like humidity, wind, dew point computing, rain-fog, rain-thunderstorm, thunderstorm, rain-tornado, tornado, fog-rain-thunderstorm, rain-thunderstorm-tornado, etc., in the context of Bangladesh.

References

1. Akter, M., Uddin, M.S., Haque, A.: Diagnosis and management of diabetes mellitus through a knowledge-based system. In: Lim, C.T., Goh, J.C.H. (eds.) 13th International Conference on Biomedical Engineering. IFMBE Proceedings, vol. 23. Springer, Heidelberg (2009).
2. Khan, M.M.H., Bryceson, I., Kolivras, K.N., Faruque, F., Rahman, M.M., Haque, U.: Natural disasters and land-use/land-cover change in the southwest coastal areas of Bangladesh. Reg. Environ. Change **15**(2), 241–250 (2015)
3. Roy, R., Chan, N.W.: An assessment of agricultural sustainability indicators in Bangladesh: review and synthesis. Environmentalist **32**(1), 99–110 (2012)
4. Rahman, M.A., Yunsheng, L., Sultana, N.: Analysis and prediction of rainfall trends over Bangladesh using Mann-Kendall, Spearman's rho tests and ARIMA model. Meteorol. Atmos. Phys. **129**(4), 409–424 (2017)
5. Bharati, S., Podder, P., Mondal, M.R.H.: Visualization and prediction of energy consumption in smart homes. Int. J. Hybrid Intell. Syst. **16**(2), 81–97 (2020)
6. Coddington, O., Lean, J.L., Pilewskie, P., Snow, M., Lindholm, D.: A solar irradiance climate data record. Bull. Am. Meteor. Soc. **97**(7), 1265–1282 (2016)
7. Olaiya, F., Adeyemo, A.B.: Application of data mining techniques in weather prediction and climate change studies. Int. J. Inf. Eng. Electron. Bus. **4**(1), 51 (2012)
8. Bharati, S., Rahman, M.A., Mondal, R., Podder, P., Alvi, A.A., Mahmood, A.: Prediction of energy consumed by home appliances with the visualization of plot analysis applying different classification algorithm. In: Frontiers in Intelligent Computing: Theory and Applications. Advances in Intelligent Systems and Computing, vol. 1014. Springer, Singapore (2020).
9. Delle Monache, L., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K.: Probabilistic weather prediction with an analog ensemble. Mon. Weather Rev. **141**(10), 3498–3516 (2013)
10. Bharati, S., Rahman, M.A., Podder, P., Robel, M.R.A., Gandhi, N.: Comparative performance analysis of neural network base training algorithm and neuro-fuzzy system with SOM for the purpose of prediction of the features of superconductors. In International Conference on Intelligent Systems Design and Applications, pp. 69–79. Springer, Cham (2019)
11. Lima, C.H., Lall, U.: Spatial scaling in a changing climate: a hierarchical Bayesian model for non-stationary multi-site annual maximum and monthly streamflow. J. Hydrol. **383**(3–4), 307–318 (2010)

12. Wu, J., Chen, E.: A novel nonparametric regression ensemble for rainfall forecasting using particle swarm optimization technique coupled with artificial neural network. In: International Symposium on Neural Networks, pp. 49–58. Springer, Heidelberg (2009).
13. Nishe, S.A., Tahrin, T.A., Kamal, N., Shahinul Hoque, M.D., Hasan, K.T.: Micro-level meteorological data sourcing for accurate weather prediction. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 353–356. IEEE (2017)
14. Mahabub, A.: A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Appl. Sci. **1**(12), 1667 (2019)
15. Bharati, S., Podder, P., Mondal, M.R.H.: Diagnosis of polycystic ovary syndrome using machine learning algorithms. In 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1486–1489. IEEE, June 2020
16. Mahabub, A., Mahmud, M.I., Hossain, M.F.: A robust system for message filtering using an ensemble machine learning supervised approach. ICIC Express Lett. Part B Appl. **10**(9), 805–812 (2019)
17. Mahabub, A.: A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. SN Appl. Sci. **2**(4), 1–9 (2020)
18. Andrade, J.R., Bessa, R.J.: Improving renewable energy forecasting with a grid of numerical weather predictions. IEEE Trans. Sustain. Energy **8**(4), 1571–1580 (2017)
19. Raihan-Al-Masud, M., Mondal, M.R.H.: Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. PLoS ONE **15**(2), e0228422 (2020)
20. Bharati, S., Podder, P., Mondal, M.R.H.: Hybrid deep learning for detecting lung diseases from X-ray images. Inf. Med. Unlocked **20**, 100391 (2020)
21. Mondal, M.R.H., Bharati, S., Podder, P., Podder, P.: Data analytics for novel coronavirus disease. Inf. Med. Unlocked **20**, 100374 (2020)
22. Podder, P., Khamparia, A., Mondal, M.R.H., Rahman, M.A., Bharati, S.: Forecasting the spread of COVID-19 and ICU requirements. Preprints (2021). 2021030447. https://doi.org/10.20944/preprints202103.0447.v1
23. Podder, P., Bharati, S., Hossain Mondal, M.: 10 Automated gastric cancer detection and classification using machine learning. In: Gupta, D., Kose, U., Le Nguyen, B., Bhattacharyya, S. (ed.) Artificial Intelligence for Data-Driven Medical Diagnosis, pp. 207–224. De Gruyter, Berlin, Boston (2021). https://doi.org/10.1515/9783110668322-010