Data Science Pre-press (2019) 1–36
IOS Press

An Overview of Weather Forecasting for Bangladesh Using Machine Learning Techniques

Atik Mahabub [a,c,*]
Al-Zadid Sultan Bin Habib [b,c]
[a] *Dept. of ECE, Concordia University, Montréal, QC H3G 1M8, Canada*
E-mail: atikmahabub1209042@gmail.com
[b] *Dept. of CSE, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh*
E-mail: al.zadid.habib@gmail.com
[c] *Dept. of ECE, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh*
E-mails: atikmahabub1209042@gmail.com, al.zadid.habib@gmail.com

Abstract. Weather forecasting bears significant impacts in our day-to-day life in every aspect from agricultural perspectives to event management. Weather forecasting becomes an uphill task for countries like Bangladesh where plain lands similarly coincide with coastal areas or hill tract areas and weather changes frequently. Weather forecasting contains some predictions of key parameters like wind speed, humidity, temperature, and rainfall. Several previous weather forecasting models used the complicated mathematical instruments which were rarely accurate. In this paper, regression-based Machine Learning (ML) models have been presented to predict the weather parameters accurately for Bangladesh. A practical application of ML techniques towards environmental numerical modeling has been developed. The raw dataset has been collected from Bangladesh Meteorological Division (BMD) which includes data of wind speed, humidity, temperature and rainfall for the past five years of Bangladesh of several weather stations across the country. Several regression algorithms have been used e.g. 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). The output of regression techniques has been compared with the existing forecast-based models which show that ML-based models are more accurate than conventional methods.

Keywords: Machine Learning, Weather Forecasting, Data Mining, Bangladesh Meteorological Division (BMD), Regression

1. Introduction

Weather forecasting is supposed to be a prime factor for Bangladesh's economy as agriculture plays a vital role in country's overall Gross Domestic Product (GDP) which accounts for approximately 20% of the total amount. Nearly 70% of its total population live in rural areas and 60% of them earn their livelihood from tillage stuff. Even a significant amount of total annual exports are from farming products which tend to be in the region of 13-18% of the country's total GDP. 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 [1–4].

*Corresponding author. E-mail: atikmahabub1209042@gmail.com.

Since the dawn of technological advancement, weather forecasting is that much noteworthy which have been endeavored to predict accurately as much as possible for many years by the experts. Initially, meteorologists collect quantitative data to assemble the forecast. Apart from its significant importance in agriculture and economy, a successful forecast of rainfall and thunderstorm can save airlines from falling into accidents or unwanted crashes which may cause the death of many people.

Moreover, accurate weather forecast can be beneficial to save times in case of unwanted flight delays. Temperature and humidity are also key metrics to influence the agriculture and the economy of the country [5–9]. A few types of data sources like land-based stations, marine, radar, weather balloons, satellite 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 [10].

Intelligent weather prediction techniques can help us to a certain degree that can help us to make effective decisions which 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, airplanes, 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 is delineated as numerical or computational climate forecasting [11–13]. Numerical and computational models of weather forecasting methods have replaced the conventional forecasting methods. Day by day, their usage has become more accustomed to atmosphere estimation [14, 15]. For example, Bayesian Networks [16] along with contrasting scaling attributes can be utilized to assess whether there remains any notable pattern in climate information.

Despite having several weather forecasting techniques, complex physics behind weather doesn't 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 [17]. 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 [18–21]. In most cases, ANN showed decent performance in weather forecasting. In this paper, several regression-based ML algorithms were implemented to predict the weather and analyze the comparative performance of those algorithms. In the case of weather parameters, we used four weather parameters e.g. wind speed, humidity, rainfall, and temperature. We used multiples statistical regression techniques like SVR, Linear Regression, Bayesian Ridge, GB, XGBoost, CatBoost, AdaBoost, KNN and DTR to run the weather forecasting [22–30]. In the end, we compared our measured output with the result of the conventional prediction method. Point to be noted that our proposed regression-based model provides better computational performance both in training and testing steps. To be candid to declare that ours is the very first regression-based weather forecasting model for Bangladesh in which all of these regression-based techniques were applied on the weather dataset provided by BMD.

Over the last few decades, researchers conducted a huge number of a research project for automated weather forecasting based on Artificial intelligence (AI) which included methods from an expert system, fuzzy logic, machine learning, data mining, deep learning, etc. Space-Time model was denoted by Tae-wong et al. [31] which visualized the short time and geographical conditions of the day by day rain

event. In case of air temperature prediction different research works were done where Chevalier et al. [32] did it using SVR. ANN-based temperature forecasting model was developed by Devi et al. [33] using real-time quantitative data regarding the ongoing state of the atmosphere. Olaiya and Adeyemo [34] likewise examined the execution of ANN and decision trees amid the grouping of most extreme, least, and mean temperature, precipitation, dissipation and wind speed on meteorological information accumulated from Nigeria. Luminto et al. [35] developed a weather forecasting model for Indonesia using Linear Regression model for predicting rice cultivation time. Navin et al. [36] created site-specific prediction models for solar power generation from weather forecasts using ML techniques. They compared numerous regression strategies for producing prediction models, including linear least squares and support vector machines utilizing multiple kernel functions. Nishe et al. [37] developed a system that collected data for accurate weather updates and predictions. They developed an Android application and a device build with Arduino and GPS system which received micro-level data from thousands of users. Zaman et al. [38] worked on ML-based rainfall prediction system for Bangladesh and he used multiple regression algorithms and showed their comparative performance. Nasimul et al. [39] developed an ML-based weather analysis model on Los Angeles weather dataset where they used the C4.5 learning algorithm. They classified different weather events such as normal, rain and fog and applied this C4.5 classification technique. Mohammadi et al. [40] anticipated the dew point temperature on the day by day scale on various atmosphere conditions applying extreme machine learning algorithms on five basic atmosphere related highlights, for example, mean air temperature, relative mugginess, environmental weight, vapor weight, and horizontal global solar radiation. Nasimul et al. [41] developed an SVR based rainfall prediction model for Bangladesh. His proposed technique outperformed the conventional method in case of accuracy level for weather data of Bangladesh. Krasnopolsky et al. [42] derived a practical application of Neural Network (NN) techniques to numerical weather modeling and prediction. They presented their model using NN as a statistical or ML method to develop vastly accurate and fast emulations for time-consuming model physics components. Ayesha Atta et al. [43] provided a mechanism to predict the traffic congestion using Artificial Neural Network (ANN) to control or mitigate the blockage for smoothening the road traffic. They used the Back Propagation algorithm (BPN) for road traffic to increase the transparency, availability, and efficiency in service to enhance the comfort level of the travelers. The prediction of congestion was operationalized by using BPN to train the NN in their research work. In [44] Ayesha Atta et al. proposed an intelligent traffic congestion control system using the Internet of Things (IoT) and Fuzzy Logic System (FLS). They utilized RFID readers to sort out the traffic congestion and the blockage at any intersection of the street. Anwar Saeed et al. proposed an optimal utilization of cloud resources using BPN and multi-level priority queue in [45]. A hybrid approach has been proposed with an objective to achieve the maximum resource utilization in cloud computing. They showed the result in simulation-based on the form of MSE and regression with job dataset comparing three distinct algorithms Scaled Conjugate Gradient (SCG), Levenberg Marquardt (LM) and Bayesian Regularization (BR). Areej Fatima et al. [46] proposed an IoT based Plant Factors of Smart City (PFSC) using Multilayer Fuzzy Logic System (MFLS). They categorized the PFSC into two levels and proposed an MFLS based Expert System (ES) to categorize the evaluation level of planet factors of the smart city into low, satisfied or good.

Considering all the above mentioned previous works we focused our research towards a definitive goal. Most of the previous research works were based on weather forecasting but not all of them had the main goal to forecast the weather only. Moreover, they used NN or other conventional regression techniques but didn't achieve the anticipated results. In contrast, all of the common regression techniques were used along with some very recent ones. Those algorithms have been applied on the weather dataset collected

from BMD and actions have been taken to observe the comparisons amongst them to forecast the weather correctly considering four weather parameters e.g. wind speed, humidity, rainfall and temperature (High & Low). In this research, we are able to forecast advanced one-year weather parameters of Bangladesh and compare it with the actual BMD datasets.

2. Methodological Formulation

The methodological formulation can be divided into three sub sections, statistical formulations, algorithm analysis and model formulation & dataset preprocessing.

2.1. Statistical Formulation

To evaluate the performance of the model, some statistical terms were considered. In order to prove the validity of our model, it has to be proven that the error percentage is low compared to its prediction accuracy. The following observations are considered:

2.1.1. Mean Absolute Error

The Mean Absolute Error (MAE) is examined to describe the average model-performance error. MAE is a more natural estimation of average error, and it's not quite ambiguous. Each of these measures is dimensioned in that sense that it represents an average model-prediction error in the units of the variable of interest. These estimations have been used to represent the average difference rather than an average error. This is done when no set of measures is acknowledged to be the most reliable [47]. The calculations of MAE is comparatively simple. It involves summing the absolute error of the errors obtained in the total error and then dividing the total error by n. To simplify, it is assumed that there are n samples of model errors $\in$ measured as $(e_i, i = 1, 2, 3 \ldots n)$. The uncertainties brought in by observation errors or the procedure to compare model and observations are not considered here. The error sample rate $\in$ is unbiased. It is assumed that $x = x_i | i = 1, 2, \ldots\ldots, n$ and $y = y_i | i = 1, 2, \ldots\ldots, n$ are two finite length discrete samples where n is the number of total sample and x_i and y_i are the values of the i^{th} sample in x and y, respectively. The MAE can be calculated for the data set as follows [48]:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \tag{1}$$

2.1.2. Mean Squared Error

The Mean Squared Error (MSE) has been one of the dominant quantitative performance evaluation metrics in the field engineering. Its ubiquitous design makes it preferable to choose to optimize the performance of the algorithms. Although this metric was used to evaluate the performance of our model, it is widely used in evaluating the work rate of signal processing algorithms. It provides a quantitative score that describes the degree of fidelity or the level of distortion between them. It is assumed that one sample is pristine whether the other one is not. We assume that $x = x_i | i = 1, 2, \ldots\ldots, n$ and $y = y_i | i = 1, 2, \ldots\ldots, n$ are two finite length discrete samples where n is the number of total sample and x_i and y_i are the values of the i^{th} sample in x and y, respectively. The MSE between these data is as follows [49]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|(x_i - y_i)|^2 \tag{2}$$

2.1.3. Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) is a measure of the efficacy of regression models. The optimal MAPE model can be proved and the universal consistency of empirical risk minimization can be shown based on the MAPE. The best model under the MAPE is almost equivalent to doing weighted MAE regression. Basically, classical regression models are acquired by choosing a model that minimizes an empirical estimation of the MSE. Due to robustness MAPE is chosen along with MAE for being regression quality measurement metric. It is assumed that $x = x_i|i = 1, 2, \ldots\ldots, n$ and $y = y_i|i = 1, 2, \ldots\ldots, n$ are two finite length discrete samples where n is the number of total sample and x_i and y_i are the values of the i_th sample in x and y, respectively. Where, n is the total number of samples or data. It can be illustrated that the formula of MAPE is given as follows [50]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right| \tag{3}$$

2.2. Algorithm Analysis

Basically, regression is a statistical approach to find the relationship between variables. In ML this is used to predict the consequence of an event depending on the relationship between variables acquired from the dataset. It is termed as one of the simplest supervised ML algorithms. These regression algorithms are used to predict the response variable set off explanatory variables. On the other hand, the term 'Boosting' is used to refer to a family of algorithms which can convert weak learners to strong learners. To convert the weak learners to strong learners, the predictions of each weak learners are combined using the methods like an average/weighted average of the prediction which has a higher vote.

2.2.1. KNN

k-Nearest Neighbors or KNN is a non-parametric method which is used for both classification and regression-type problems. In the case of KNN regression, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors. It is a sort of instance-based learning or lazy learning, where the functions are only predicted locally and all the computations are deferred. It is supposed to be the simplest of all ML algorithms. The KNN algorithm is used for estimating continuous variables. It uses a weighted average of the *k* nearest neighbors, weighted by the inverse of their distance. First of all, the Euclidian or Mahalanobis distance is computed from the query example to the labeled examples. Then the labeled examples are ordered by increasing distance. To the next step, it is tried to find a heuristically optimal number of *k* of nearest neighbors, based on the values of Root Mean Squared Error (RMSE). This is done using cross-validation. In the end, an inverse distance weighted average is calculated with the k-nearest multivariate neighbors [51].

2.2.2. AdaBoost

Adaptive Boosting or shortly AdaBoost is the first practical boosting algorithm. It converts a set of weak classifiers into a strong one. It can be used in conjunction with other types of learning algorithms. The output of the other algorithms is combined into a weighted sum that represents the output of the boosted algorithm. Although, AdaBoost is sensitive to noisy data and outliers. In some cases, it can be

less credulous to the overfitting problem than other learning algorithms. The final equation for AdaBoost is represented as follows [52]:

$$F(x) = sign(\sum_{m=1}^{M} \Omega_m f_m(x)) \tag{4}$$

Where f_m represents the m^{th} weak classifier and Ω_m is the corresponding weight. It is exactly the weighted combination of M weak classifiers.

2.2.3. SVR

Support Vector Regression or SVR is based on statistical learning theory. It is widely used in both classification and regression types problems. Its most efficiency is observed in forecasting in financial data and time series prediction. The SVR must use a cost function to measure the estimated risk in order to lessen the regression error. One might choose a loss function to calculate the cost from least module loss function, quadratic loss function, etc. The insensitive loss function exhibits the sparsity of the solution. It contains a fixed and symmetrical margin term. It runs into the risk of overfitting the data with poor generalization if the margin is either zero or very small. On the contrary, if the margin tends to be large, it gains a better generalization at the risk of having a higher testing error. Generally, the estimation function in SVR takes the following form [53]:

$$f(x) = (\omega.\phi(x)) + b \tag{5}$$

In Eq. 5, $(.)$ denotes the inner product in Ω, a feature space of possibly different dimensionality such that $\omega : X \rightarrow \Omega$ and $b \in R$. The other two parameters, ω and b can be determined from the training dataset by minimizing the regression risk based on the estimated risk.

2.2.4. Bayesian Ridge

Bayesian Ridge regression technique can be used to include regularization parameters in the estimation procedure. The regularization parameter is not set in a hard sense but tuned to the data at hand. This can be done by introducing uninformative priors over the hyperparameters of the model. The L_2 regularization used in ridge regression is supposed to be equivalent to finding a maximum a posteriori estimation under a Gaussian prior over the parameters ω with the precision $\lambda^{-}1$. Instead of setting lambda manually, it's treated as a random variable to be estimated from the data. A probabilistic model of the regression problem is estimated in Bayesian Ridge. The prior for the parameter ω is given by a spherical Gaussian. The general formulation for Bayesian Ridge is done as follows [54]:

$$p(\omega|\lambda) = N(\omega|\theta, \lambda^{(} - 1)I_p) \tag{6}$$

The priors over α and λ are chosen to be gamma distributions, the conjugate prior for the precision of the Gaussian. The resulting model is called Bayesian Ridge Regression which is similar to the classical Ridge. The parameters ω, α, and λ are calculated jointly during the fit of the model. The remaining other hyperparameters are the parameters of the gamma priors over α and λ. These are usually chosen to be non-informative and are calculated by maximizing the marginal log-likelihood.

2.2.5. Linear Regression

Linear Regression is a very commonly used method for predictive analysis. It attempts to model the relationship between two variables by fitting a linear equation to observed data. Here, one variable is

considered as an explanatory variable and the other is assumed as a dependent variable. It would be first determined that there exists a relationship between the variables of interest or not. If there is no correlation between the assumed explanatory and dependent variables, then fitting a linear regression model will not bring any fruitful conclusion. The correlation coefficient is considered to be one of the most valuable numerical measures of association between two variables which resides between -1 and 1 indicating the strength of the association observed data for the two variables. This may form the following equation [55]:

$$Y = a + bX \tag{7}$$

Where X and Y correspondingly refer to the explanatory variable and the dependent variable. b is considered to be the slope of the line, and a is the intercept (the value of y when $x = 0$).

2.2.6. GB

Gradient Boosting or Gradient Tree Boosting is a generalization or further development of AdaBoost. This statistical framework cast boosting as a numerical optimization problem where the objective is to minimize the loss of the model by adding weak learners using a gradient descent like procedure. This sorts of algorithms were defined as a stage-wise additive model. This is done as one new weak learner is added at a time and existing weak learners in the model are frozen and left unchanged. This generalization allowed arbitrary differentiable loss function to be used, which strengthens the technique to support regression, multi-class classification and more. The general formula for the GB can be stated as follows [56]:

$$F_{m+1}(x) = F_m(x) + h(x) = y \tag{8}$$

Where m is each stage, F_m is some imperfect model, y is the output variable, and h is an added estimator to provide a better model.

2.2.7. XGBoost

XGBoost is one of the widely used ML algorithms. It can be used for supervised learning tasks like regression, classification, and ranking. It is developed on the principles of the GBM framework and designed to push the extreme of the computation limits of machines to provide a scalable, portable and accurate outcome. For a given data set with n examples and m features $D = (x_i, y_i)(\mid D \mid = n, x_i \in R^m, y_i \in R)$, a tree ensemble model, uses K additive functions to predict the output [57].

$$y_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{9}$$

Where $F = \{f(x) = \omega_q(x)\}(q : R^m \rightarrow T, \omega \in R^T)$ is denoted as the space of the regression trees. Apart from that, q represents the structure of each tree which maps an example to the corresponding leaf index, T is the number of leaves, f_k denotes to an independent tree structure q and leaf weights ω. Each regression tree contains a continuous score on each of the leaf, where ω_i is used to represent score on the i^{th} leaf.

2.2.8. CatBoost

CatBoost is a recently developed open-sourced ML algorithm from Yandex. It can be integrated with deep learning frameworks like Google's TensorFlow or Apple's CoreML without any hazards. It can work with different types of data and cooperates to solve a wide range of problems that we face today. It

provides the best accuracy in two ways. Firstly, it yields the state of the art results without any extensive data training which is required by other ML models. It can provide enormous support for the descriptive data formats. This works well with multiple categories of data e.g. audio, text, image, and historical data. CatBoost works better where categorical features play a vital role. But it takes up a bit longer training time than GBM while its prediction time is faster than other algorithms. Its general formula can be formed as [58]:

$$F_m(x) = F_(m-1)(x) + \gamma_m h_m(x) \tag{10}$$

$$\gamma_m = \arg min \sum_{i=1}^{n} L(y_i, F_(m-1)(x_i) + \gamma h_m(x_i)) \tag{11}$$

In Eq. 10 and 11, $F_m(x)$ is the final output, L is the loss function, m is the number of iterations, $h_m(x)$ is the pseudo-residuals, and γ_m is the multiplier

2.2.9. DTR

Regression models can be built in the form of a tree structure using DTR (Decision Tree Regression). Initially, a dataset is broken into small and smaller subsets and at the same time, an associated decision tree is developed in the incremental order which results in decision nodes or leaf nodes. Usually, a decision node has two branches, each representing values for the attribute tested. Whether the leaf node represents a decision on the numerical target. The topmost decision node in a tree which is referred to as the best predictor is called the root node. It can be stated that decision trees can handle both categorical and numerical data. The $ID3$, $C4.5$, and $CART$ algorithms are used to construct a decision tree for regression by replacing information gain with standard deviation reduction. $ID3$ involves top-down and greedy search through the space of possible branches with no backtracking [59].

2.3. Model Formulation & Dataset Preprocessing

The collected raw dataset from BMD was preprocessed and a lot of cleaning was required to convert it from semi-structured dataset to structured dataset and prepared for the implementation of our desired model. We considered the data of the time period of 2012 to 2018 as we wanted to obtain the most possible suitable output for Bangladesh for the most recent weather data. After completing the preprocessing of the dataset initially we had 2 separate datasets in CSV file, one for 2012-2017, for training and testing the model and another one for 2018 to make the forecasting where each dataset had 4 specific parameters e.g. wind speed, rainfall, humidity, and temperature (low and high). We conducted our research on this preprocessed dataset and later compared all the outcomes for better predictions. Provided data had the category as rainfall (millimeter), humidity (percentage of water in the air), wind speed (kilometer per hour), temperature (degree Celsius). The weather data was collected from 33 weather stations across Bangladesh which stations are considered the core government-controlled weather station of Bangladesh. Each data file had a similar column structure having a year, month, day and other corresponding parameter values. Basically, long-range forecasting can be divided into 4 categories, (a) periodicity approach (b) correlation approach (c) extended synoptic approach and (d) dynamical approach [60]. BMD is a government agency for weather prediction in Bangladesh. In 2007 BMD first introduced statistical forecast system based on ensemble technique [61, 62]. Although their predictions were acceptable that was always dependent on some specific predictors. However, we tried to put it one-step forward through our research. After preprocessing and dataset cleaning to make it structural

dataset, the raw dataset was split into two parts, one for training data and another one for testing data. The model would be trained up using these regression-based learning algorithms in the training dataset. Next, the performance would 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 used the first dataset file which contains the weather data for the years between 2012 and 2017. We used the second dataset file containing weather data of 2018 to forecast the weather and match it with collected data. This forecasting performance was monitored later.

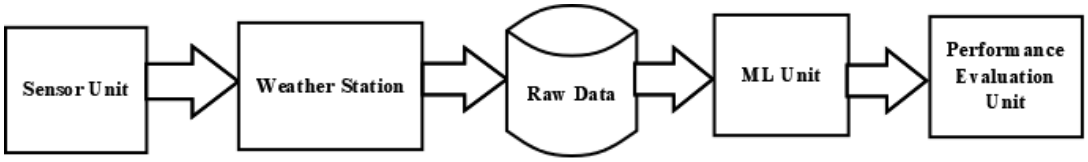

Fig. 1. ML-based weather forecasting model for weather parameters (rainfall, humidity, wind speed, low temperature, and high temperature) for Bangladesh.

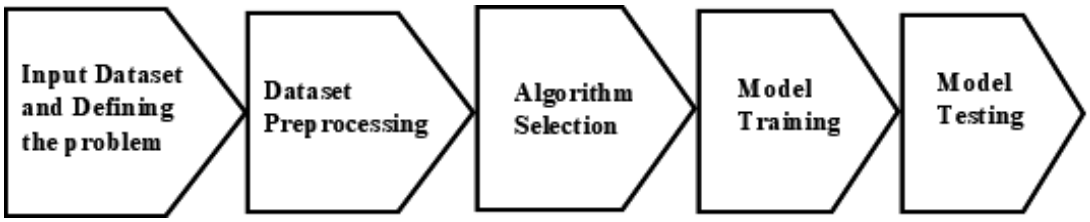

Fig. 2. The workflow of the ML Unit.

Figure 1 illustrates the block diagram for the weather forecasting model of Bangladesh using ML-based techniques. Here, several blocks play different roles. The sensor unit usually consists of different sensors (a.k.a. humidity sensor, temperature sensor, wind speed detector, rainfall detector, etc.) to evaluate the initial values of weather parameters over a specific period of time. Usually, these data are collected via weather stations established in several places across the country. The raw data is collected and stored in the database. The ML unit plays its role in raw data and finally provides the output. Multiple ML algorithms (a.k.a. KNN, XGBoost, GB, DTR, AdaBoost, SVR, Linear Regression, Bayesian Ridge, and CatBoost) have been applied separately as separate models. Specific models perform these specified actions according to the structure of the algorithm and then provides the output to the next unit. They have been selected to build specific models of each algorithm and combined together to form a unique ML-based model for weather forecasting. These algorithms have been selected as they show better performance in case of analyzing weather data. In the end, the final output can be monitored by the Performance Evaluation Unit. Several attributes have been defined to evaluate the performance (e.g. prediction, MAE, MSE, and MAPE). Based on the performance of each algorithm the best algorithm has been selected for each problem or each weather forecasting parameter. Figure 2 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 specified weather parameters. Here, at first, we have decided what to evaluate from these data. Initially, four weather parameters have been selected to

forecast 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, the model shows better performance with training data while being trained up. On the other hand, 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 needs to be changed or the algorithms should be changed or more data may be required or columns in the dataset should be selected properly for the evaluation.

3. Performance Evaluation

Performance Evaluation can be divided into two categories e.g. weather parameter analysis and forecasting analysis.

3.1. Weather Parameter Analysis

The embraced metrics of weather dataset covers the following information about weather and climate:

3.1.1. Wind Speed

The air movement at a particular point is shown by this parameter. It gives an idea with respect to the situation of low and high-pressure regions. In the occasion that breeze speed expands, it is an indication of fortification of pressure systems. It brings cool or hot air from the spot of root or from the spots through which they pass. By definition, Wind speed, or wind stream speed, is an essential barometrical amount brought about via air moving from high to low weight, normally because of changes in temperature. Note that breeze bearing is normally practically parallel to isobars (and not opposite, as one may expect), because of Earth's turn. Wind speed influences climate anticipating, aeronautics, and oceanic activities, development ventures, development and digestion rate of many plant species, and endless different ramifications [63, 64]. Wind speed is presently ordinarily estimated with an anemometer, however, can likewise be arranged to utilize the more seasoned Beaufort scale, which depends on close to the home perception of explicitly characterized breeze impacts [65]. From the given climate dataset, we can incorporate the yearly, monthly and daily average of wind speed for 2012 to 2017 which is represented in Figure 3, Figure 4 and Figure 5. From Figure 3, it can be seen that in 2015, the average wind speed is a little bit of low and peaks the highest value in 2017. The wind speed is high between the months from April to July but rapidly decreases afterward. The wind speed at its lowest value between the months from October to December. It can be observed from Figure 4. From the observation from Figure 5, it can be seen that wind speed is highest at the later stage of a month.

3.1.2. Rainfall

Rainfall is a key element of climate and plays a vital role in weather forecasting. Apart from that, it affects the ecosystem heavily and the agriculture-based economy is vastly dependent on rainfall. As per definition, rain is fluid water as beads that have consolidated from climatic water vapor and afterward turned out to be sufficiently substantial to fall under gravity. Rain is a noteworthy segment of the

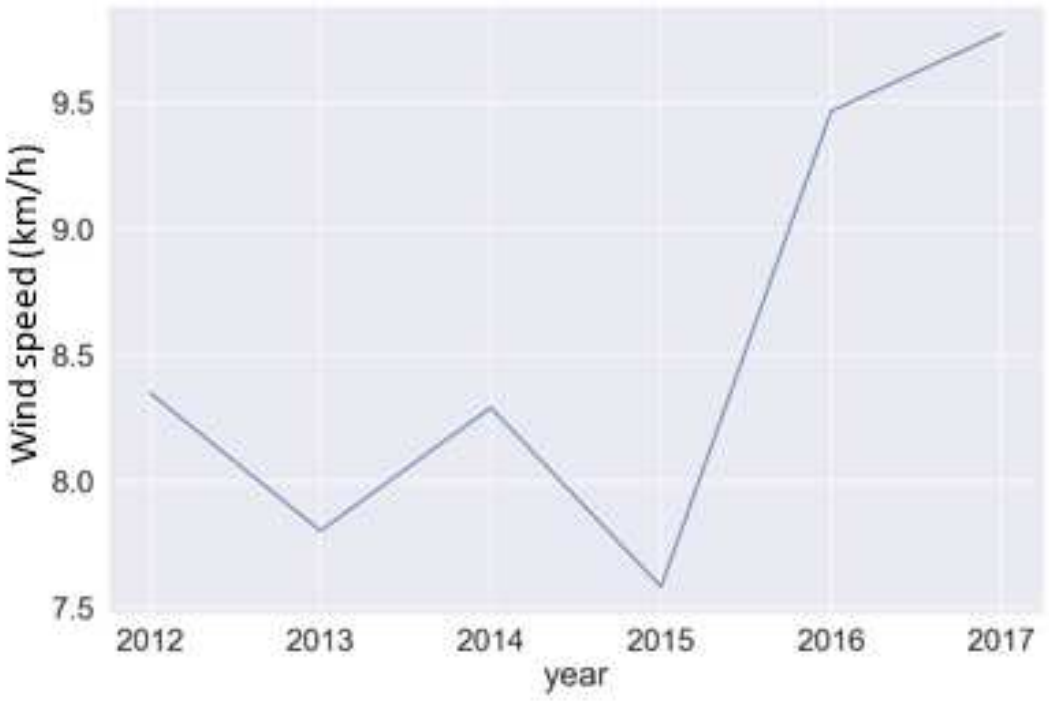

Fig. 3. Yearly average wind speed for the given dataset.

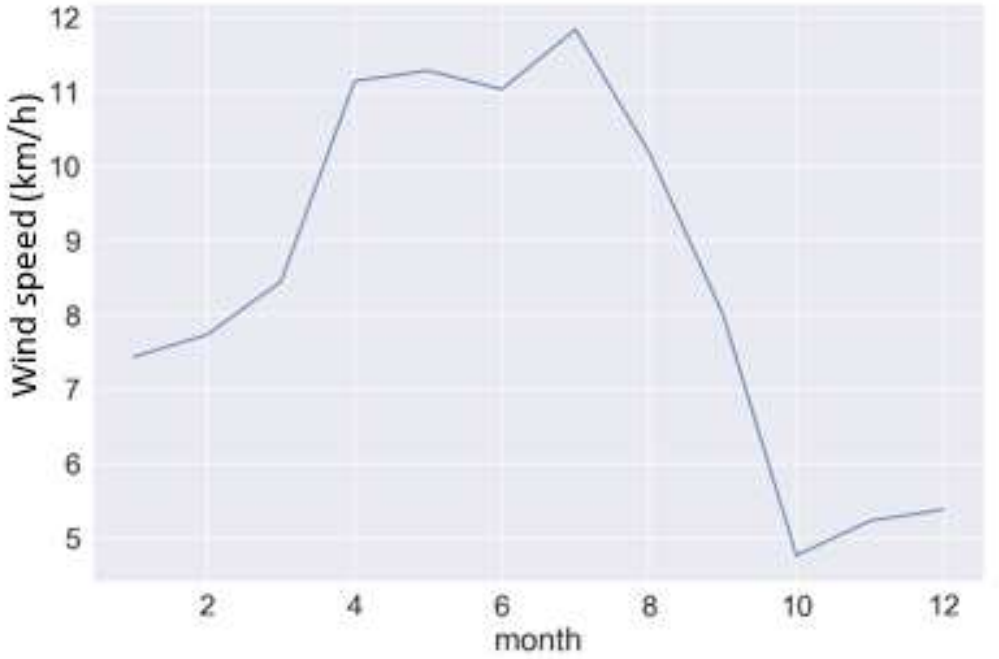

Fig. 4. Monthly average wind speed for the given dataset.

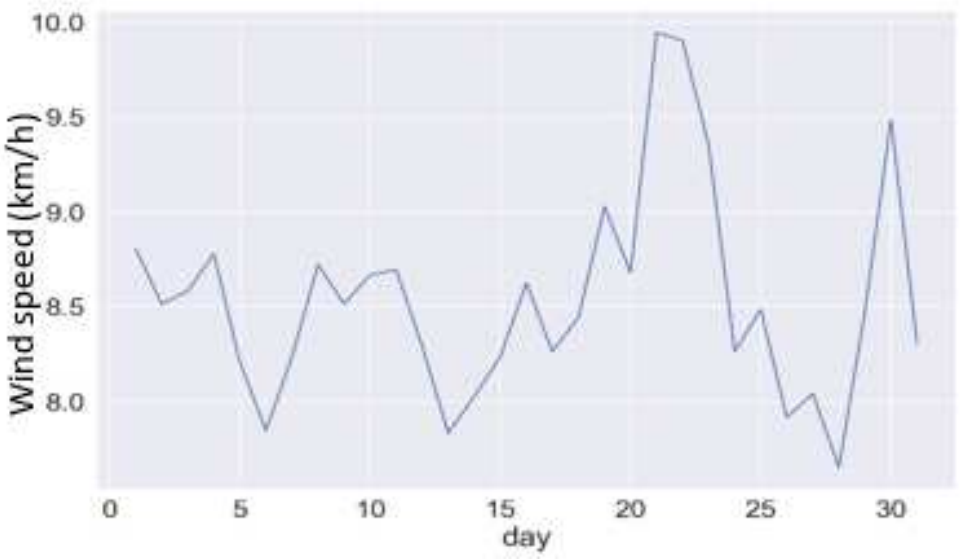

Fig. 5. Daily average wind speed for the given dataset.

water cycle and is in charge of storing the greater part of the new water on the Earth. It gives reasonable conditions to numerous sorts of biological systems, just as water for hydroelectric power plants and yield water system. The significant reason for rain creation is dampness moving along three-dimensional zones of temperature and dampness contrasts known as climate fronts. In the event that enough dampness and upward movement are available, precipitation tumbles from convective mists (those with solid upward vertical movement, for example, cumulonimbus (thunder mists) which can sort out into tight rain bands. In rugged regions, overwhelming precipitation is conceivable where the upslope stream is expanded inside windward sides of the territory at rising which powers soggy air to gather and drop out as precipitation at the edges of mountains. On the leeward side of mountains, desert atmospheres can exist because of the dry air brought about by downslope stream which causes warming and drying of the air mass. The development of the rainstorm trough, or intertropical union zone, conveys stormy seasons to savannah climes [66–68]. From our dataset, we can calculate the yearly, monthly and daily average of rainfall of Bangladesh. Here, Figure 6, Figure 7 and Figure 8 respectively correspond to the yearly, monthly and daily average rainfall across the country during the time period of 2012 to 2017. The least amount of rainfall was in 2016 which is found from Figure 6 and the amount is approximately 1.75 mm. On the other hand, the highest amount of rainfall was nearly less than 3.50 mm which was recorded in 2015 and that also can be depicted from Figure 6. It can be noticed from Figure 7 that highest amount of average monthly rainfall is recorded from May to July. Although, it slowly starts to decrease after June and obtains the highest peak exactly in June. The least amount of monthly average rainfall is recorded in two several time frames e.g. January to February and November to December. In the end, it can be illustrated from Figure 8 that the highest peak of daily average rainfall is achieved in the first five later and the lowest peak is achieved in the later five days that means 6^{th} to 10^{th} day of the month.

3.1.3. Humidity

Humidity is the measure of water vapor present noticeable all around. Water vapor, the vaporous condition of water, is commonly imperceptible to the human eye. Moistness shows the probability for

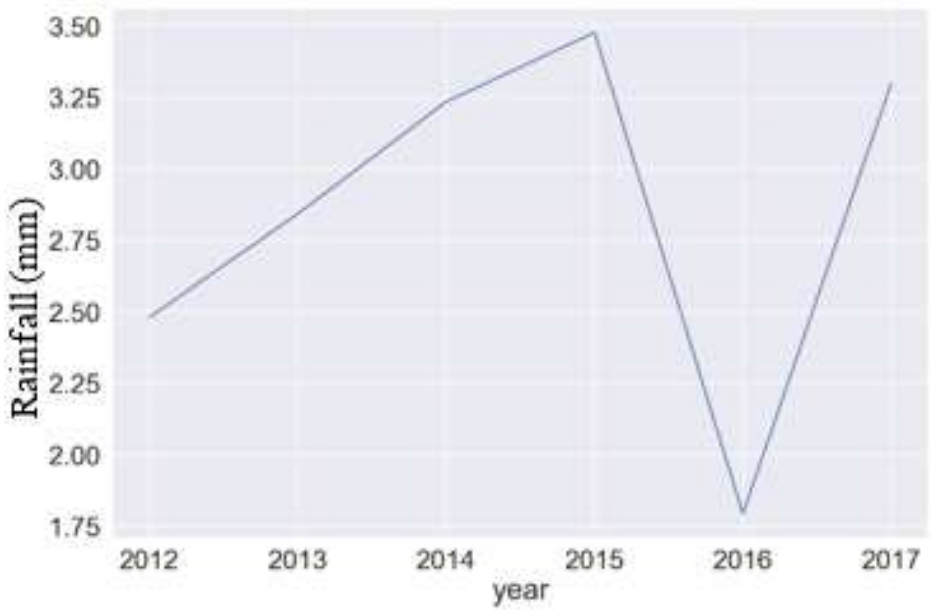

Fig. 6. Yearly average of rainfall for the given dataset.

precipitation, dew, or haze to be available. The measure of water vapor expected to accomplish immersion increments as the temperature increments. As the temperature of a package of air diminishes it will, in the end, achieve the immersion point without including or losing water mass. The measure of water vapor contained inside a package of air can fluctuate fundamentally. Existing water at different ratios in the water vapor in our climate is defined as humidity. The surface temperature would be a lot less than present ones if there would be no water vapor in the environment [69]. Three essential estimations of humidity are generally utilized: total, relative and explicit. Total mugginess portrays the water substance of the air and is communicated in either gram per cubic meter or grams per kilogram. Relative mugginess, communicated as a rate, demonstrates a current situation with total stickiness in respect to most extreme dampness given a similar temperature. Explicit mugginess is the proportion of water vapor mass to add up to sodden air bundle mass [70, 71]. Yearly, a monthly and daily average of humidity can be calculated from our given dataset which is depicted in Figure 9, Figure 10 and Figure 11. These figures are generated based on the given dataset for the years between 2012 and 2017. It can be stated in Figure 9 that highest 77% humidity is obtained in the time period between 2012 and 2013 and decay is noticed afterward. The lowest peak is achieved in 2015 which is nearly 73% humidity and later a gradual increase is found. From Figure 10, it can be mentioned that the highest amount of 85% humidity is recorded in July and the lowest amount is recorded in February which is less than 65% humidity. Apart from these, it can be observed from Figure 11 that the highest amount of humidity is obtained in the last five days of months and the least amount is found in the middle of months which is the specifically 14^{th} day of the month.

3.1.4. High Temperature

Like most of the other weather parameters, high temperature also contains the distinct feature of yearly summation of high temperature. The high-temperature gauge for the present day is determined utilizing the most elevated temperature found between 7 pm that night through 7 am the next morning. In this

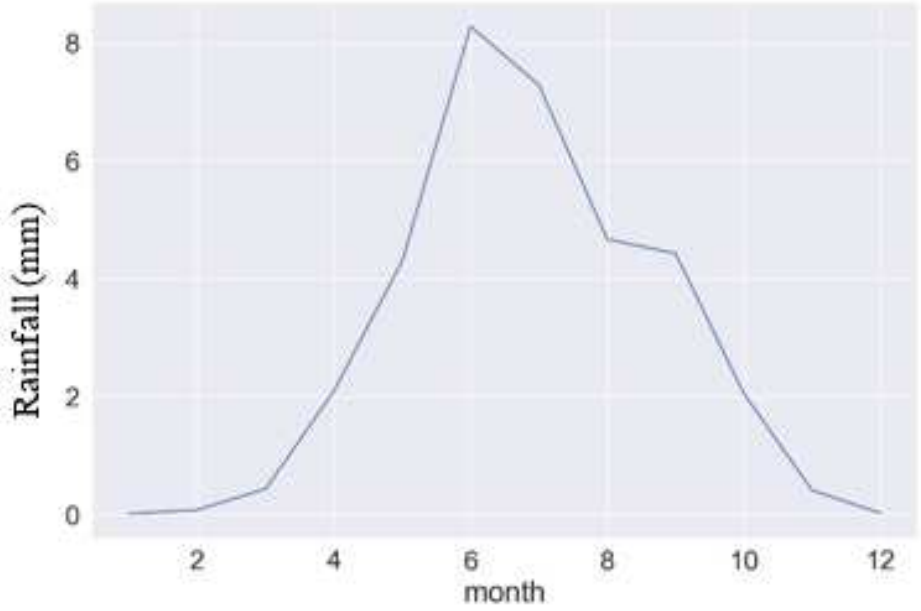

Fig. 7. Monthly average of rainfall for the given dataset.

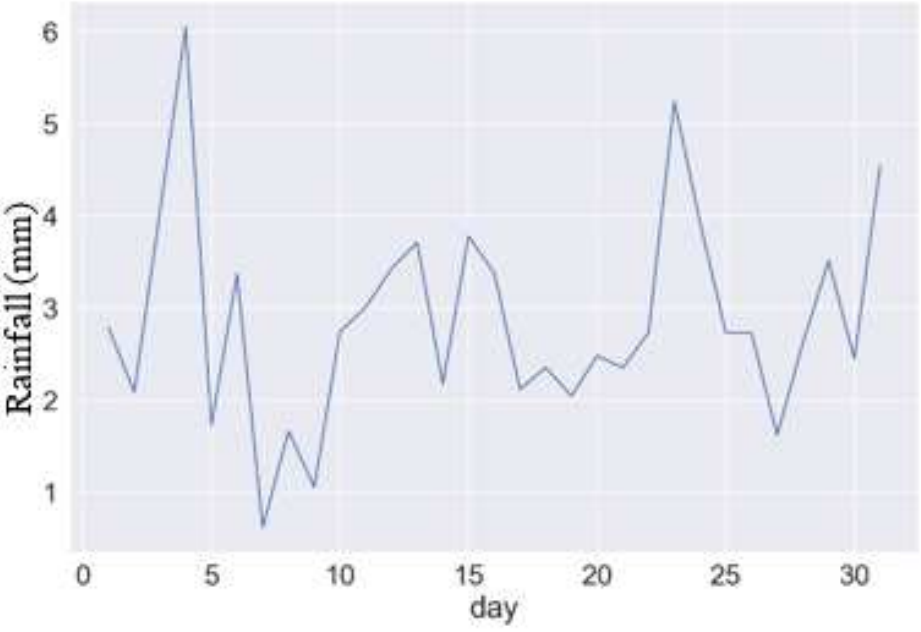

Fig. 8. Daily average of rainfall for the given dataset.

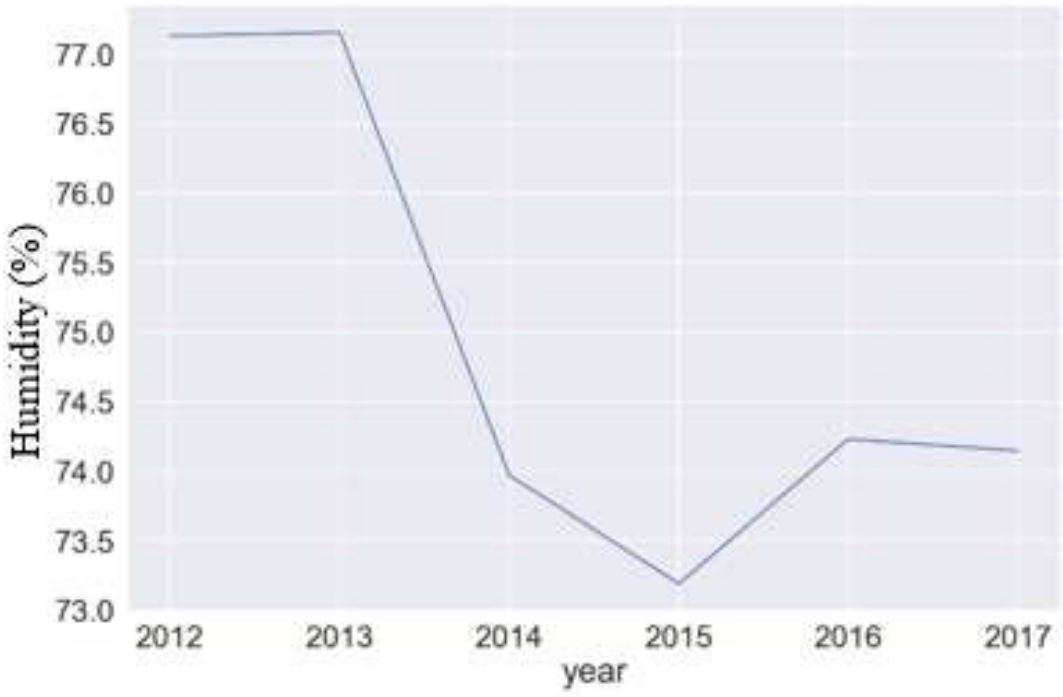

Fig. 9. Yearly average of humidity for the given dataset.

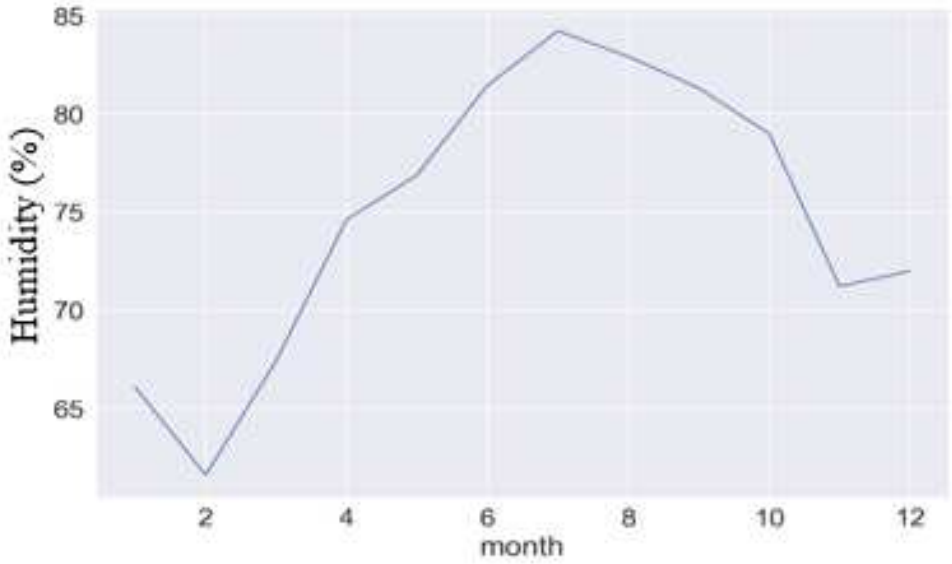

Fig. 10. Monthly average of humidity for the given dataset.

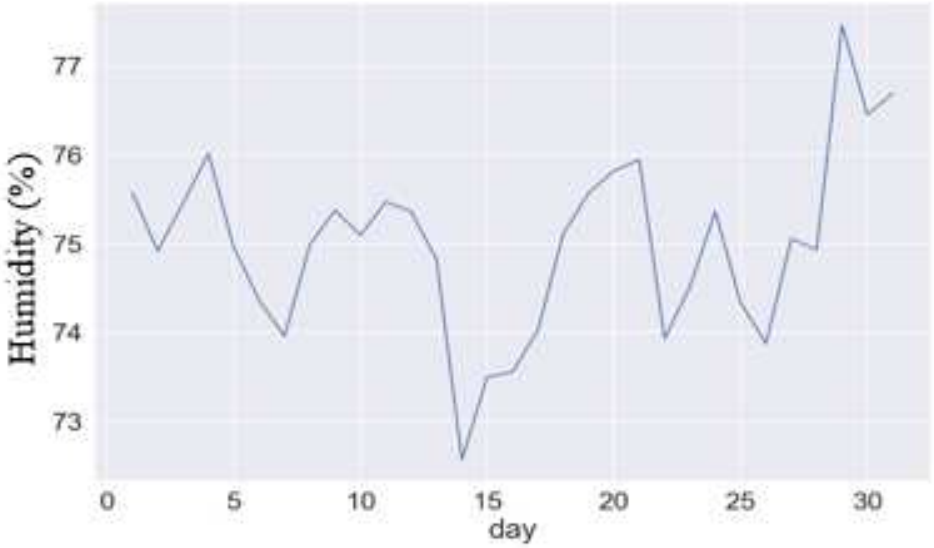

Fig. 11. Daily average of humidity for the given dataset.

way, to put it plainly, the present gauge high is in all probability tomorrow's high temperature [72, 73]. The yearly, monthly and daily average high temperature for the years 2012 to 2017 is illustrated in the following Figure 12, Figure 13 & Figure 14. From Figure 12, it can be depicted that the highest peak for high temperature is recorded for 2016 and the recorded high temperature is 30.75 degree Celsius. Apart from this, the lowest peak is found at the start of 2012 and the accumulated high temperature is 29.50 degree Celsius. It can be demonstrated in Figure 13 that the highest peak above 32 degree Celsius is achieved in May and the lowest value is obtained at the start of January which is less than 26 degree Celsius. It can be described from Figure 14 that the highest value for the daily high temperature is gained between the 10^{th} and 15^{th} day of the month. In this case, the lowest value is recorded in the later stages of the month meaning that on the 30^{th} day of the month or later.

3.1.5. Low Temperature

In case of low temperature, it does not differ from the previous one comparing to having features. Likewise, it also contains the factor: Yearly summation of low temperature. Similarly, the low temperature can also be defined in the same manner. The low-temperature measure for the present day is resolved using the most raised temperature found between 7 pm that night through 7 am the following morning. Along these lines, basically, the present gauge low is no doubt tomorrow's low temperature [74]. Figure 15, Figure 16 & Figure 17 depicts the average low-temperature correspondingly for the yearly, monthly and daily basis. The estimated measure of yearly average low temperature can be detailed from Figure 15 where the highest value of 23 degree Celsius is recorded in 2016 and the lowest one of 22.30 degree Celsius is found in 2014. From Figure 16, it can be ascribed that the highest value of 26 degree Celsius of low temperature is obtained from May to September and the lowest value is found in the start of January which value is less than 16 degree Celsius. It is found from Figure 17 that the highest value of low temperature is found on the 30^{th} day of months and the value is greater than 23 degree Celsius. On the other hand, the lowest value for low temperature is recorded just after the 25^{th} day of the month and its value is assumed to be way less than 22.40 degree Celsius.

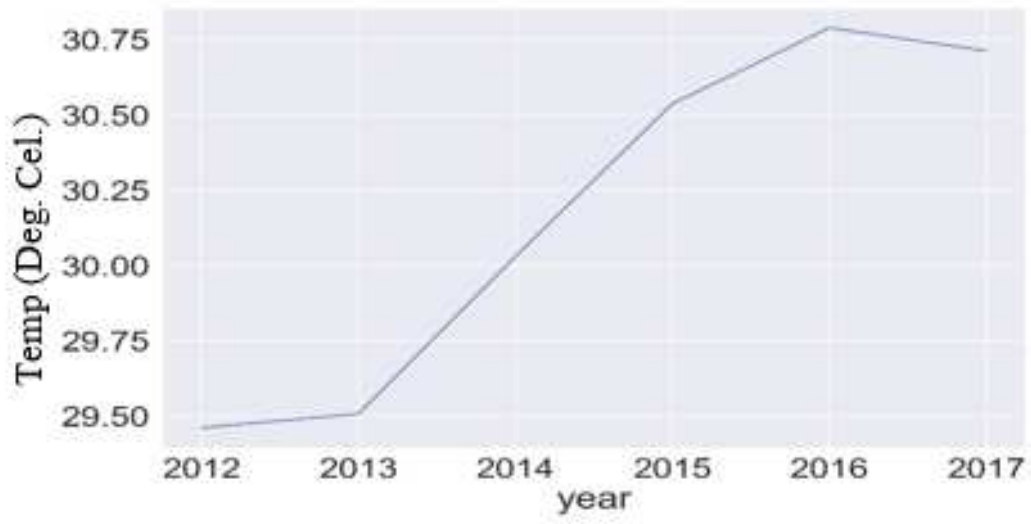

Fig. 12. Yearly average of high temperature for the given dataset.

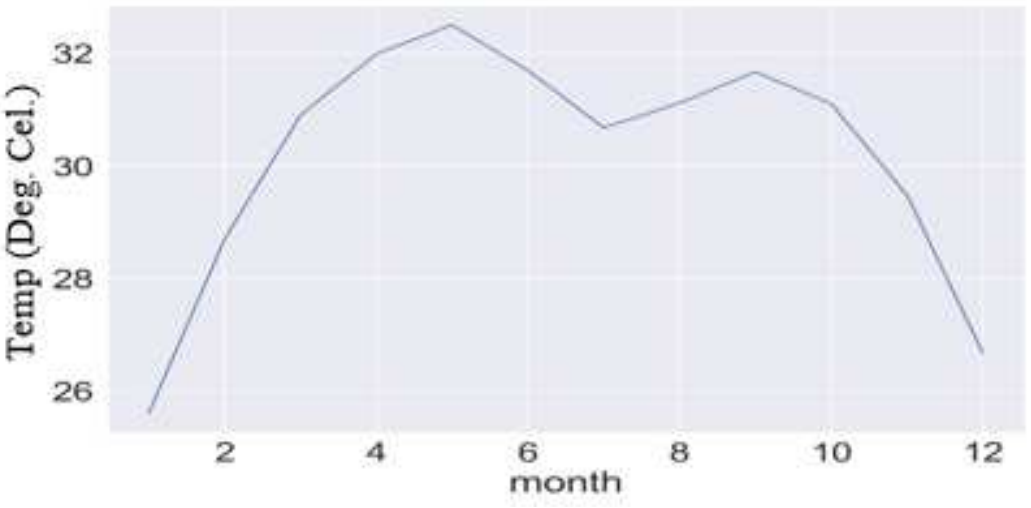

Fig. 13. Monthly average of high temperature for the given dataset.

3.2. Forecasting Analysis

Multiple regression-based ensemble algorithms were used to observe the performance of our desired model. Initially, the weather dataset is in CSV file and Python 3.6 programming language has been used for the analysis and forecasting. This dataset contained the weather data and statistics of Bangladesh for the time period during 2012-2017. This dataset was to train and test the model. After preprocessing, the dataset was split randomly into training and testing dataset to train the algorithms and to analyze their performances at random split. To demonstrate the performance of the statistical model we set three distinct factors to measure: MAE, MSE, and MAPE to check the performance of each algorithm have been used to illustrate this statistical model. Apart from that, predictions were also represented

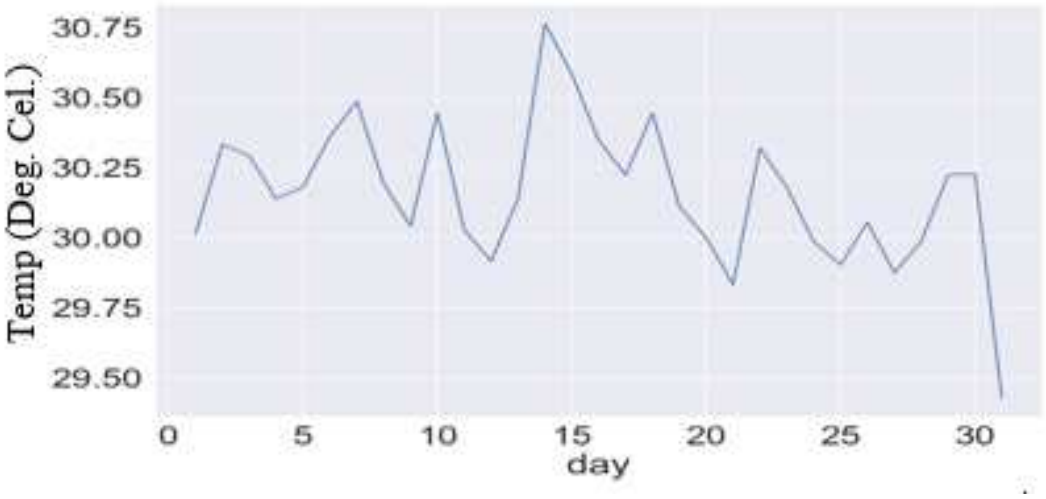

Fig. 14. Daily average of high temperature for the given dataset.

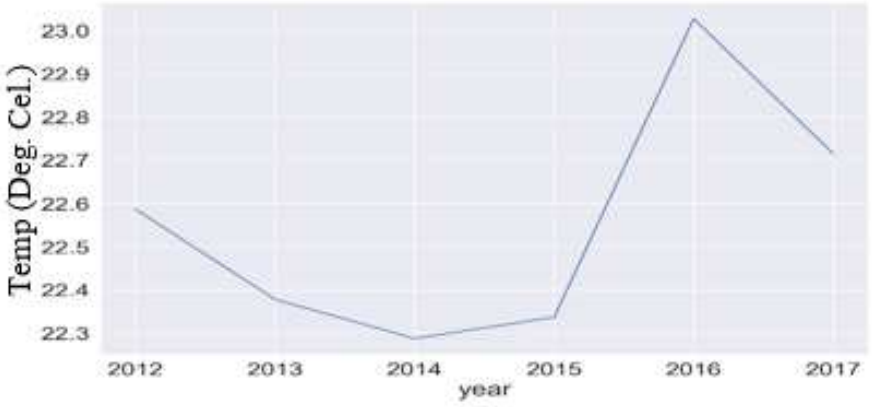

Fig. 15. Yearly average of low temperature for the given dataset.

graphically for each algorithm compared between the training data (represented as actual value) and testing data (represented by the predicted value).

From Table 1, we can analyze the statistical data for wind speed. From the MAPE values, we can identify that DTR shows more errors in predicting than other algorithms. Consequently, the MSE and MAE values are greater than others for DTR. On the contrary, CatBoost shows better performance in case of showing least error in prediction making which is verified by the MAPE, MSE and MAE values. AdaBoost and GB show similar kinds of performance respectively closer to DTR and CatBoost. Rest shows moderate performance and their statistical values lie between CatBoost to DTR.

From Figure 18, the performance of our used algorithms can be observed. Here, blue lines indicate the actual prediction which is done with training data. The red-colored predicted line shows the output for testing data. 50 observations (X-axis) were taken randomly to predict the wind speed which is plotted on the Y-axis. These curves show that our model learned well enough to predict in the test data. For each and every case every algorithm showed solid performance and predicted well in testing data compared to

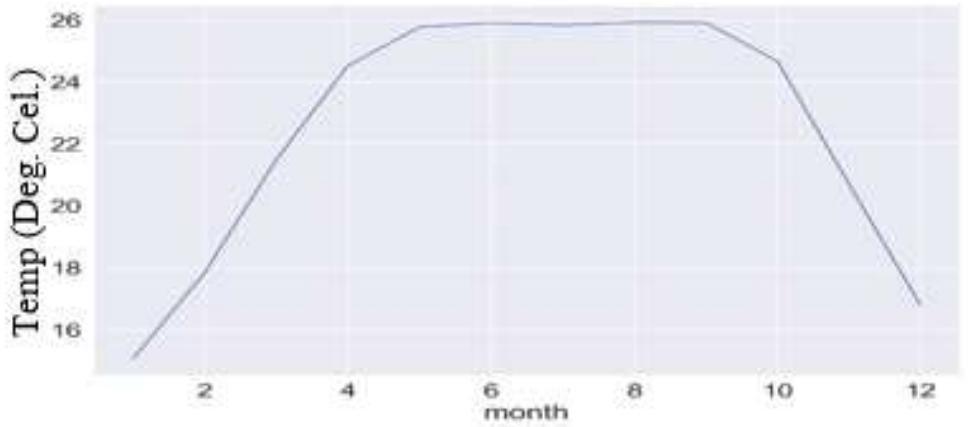

Fig. 16. Monthly average of low temperature for the given dataset.

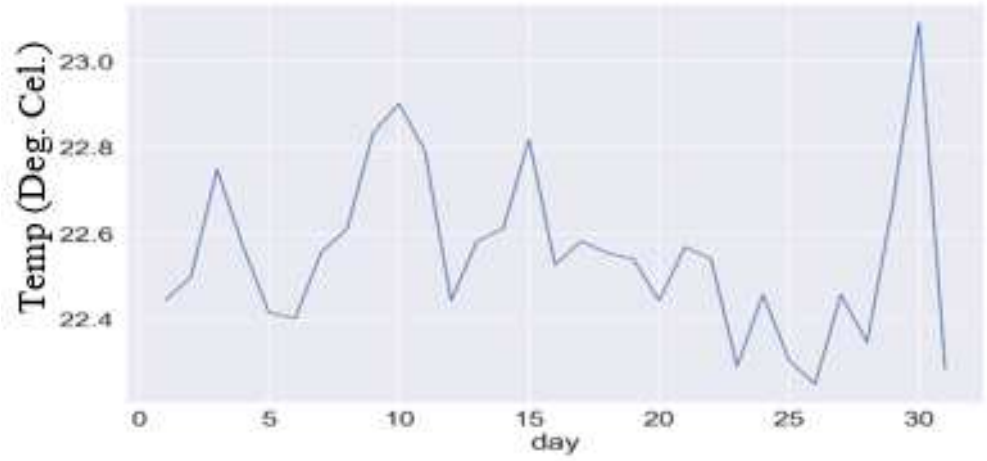

Fig. 17. Daily average of low temperature for the given dataset.

Table 1
Statistical analysis for wind speed for the 2012-2017 dataset.

Algorithms	MAE	MSE	MAPE
KNN	2.94	16.18	34.83%
XGBoost	2.79	15.48	33.02%
GB	2.78	14.91	32.85%
DTR	3.44	24.20	40.62%
AdaBoost	3.22	17.82	38.10%
SVR	3.11	17.38	36.81%
Linear Regression	3.04	16.52	35.90%
Bayesian Ridge	3.03	16.57	35.78%
CatBoost	2.71	14.65	32.04%

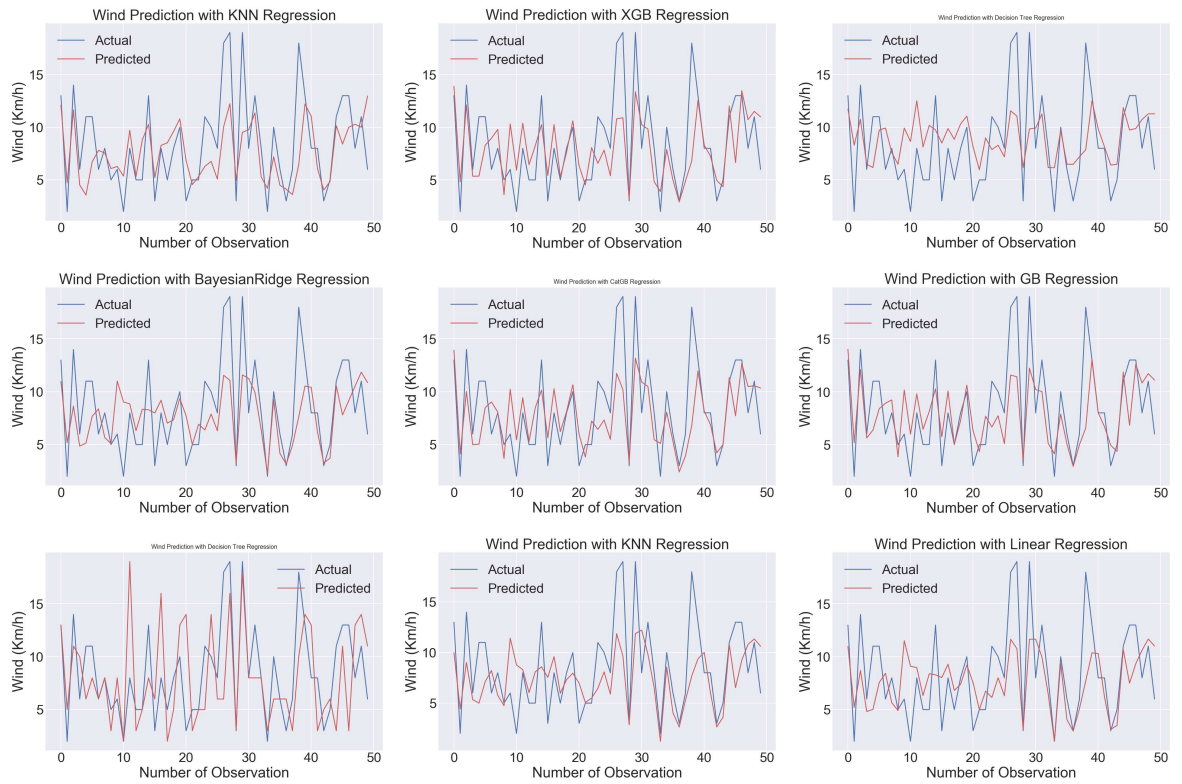

Fig. 18. Performance comparison for randomly split training data and testing data for fifty observation for wind speed prediction using nine ML algorithms(KNN, BR, LR, AdaBoost, GB, XGBoost, CatBoost, DTR, SVR).

the training data. This shows the efficacy of our chosen algorithms for our selected dataset. The model learned well and implemented the learning properly on the testing dataset. Then we considered the second dataset file which contained the weather data of 2018. This dataset will be used for forecasting with the previous algorithms. This dataset will be used for comparisons the actual values with the ML algorithms forecasted values for wind speed in 2018. The forecasting has been observed on a monthly and daily basis.

From Table 2, we can observe the statistical analysis for the forecasting by our model. A noticeable fact arises that DTR showed excellent performance for wind speed forecasting whether it was not that much fruitful for random 50 observations in Table 1. It can be defined as that the method of DTR is a quality analysis and forecasting method, most frequently utilized in technology [75]. Therefore, in the case of forecasting and having sufficient data DTR shows better performance than others do. Apart from DTR, other algorithms showed moderate performance in wind speed forecasting and their MAE, MSE and MAPE are quite similar compared to Table 1. Their error percentage values approximately fall in 28% to 36%. Similarly, we can observe the monthly and daily prediction of wind speed for each day of the year. The performance was showed comprising off all the algorithms we used later compared with the conventional value. Figure 19 and Figure 20 depicts the forecasting of wind speed for a monthly and daily basis. Figure 19, illustrates the monthly forecasting of wind speed across the country whether Figure 20 demonstrates the prediction of the daily wind speed for each day of the

Table 2
Statistical analysis for wind speed for the 2018 dataset.

Algorithms	MAE	MSE	MAPE
KNN	3.42	20.27	35.04%
XGBoost	2.92	13.62	29.89%
GB	2.90	12.99	29.66%
DTR	0.37	2.09	3.78%
AdaBoost	3.50	17.86	35.84%
SVR	3.51	21.60	35.90%
Linear Regression	3.51	20.24	35.96%
Bayesian Ridge	3.51	20.07	35.89%
CatBoost	2.82	12.97	28.82%

year. On both occasions, our proposed model forecasted quite closely to the original values. Most of the algorithms were to the point to predict accurately regarding the corresponding wind speed which verifies the stability of our model.

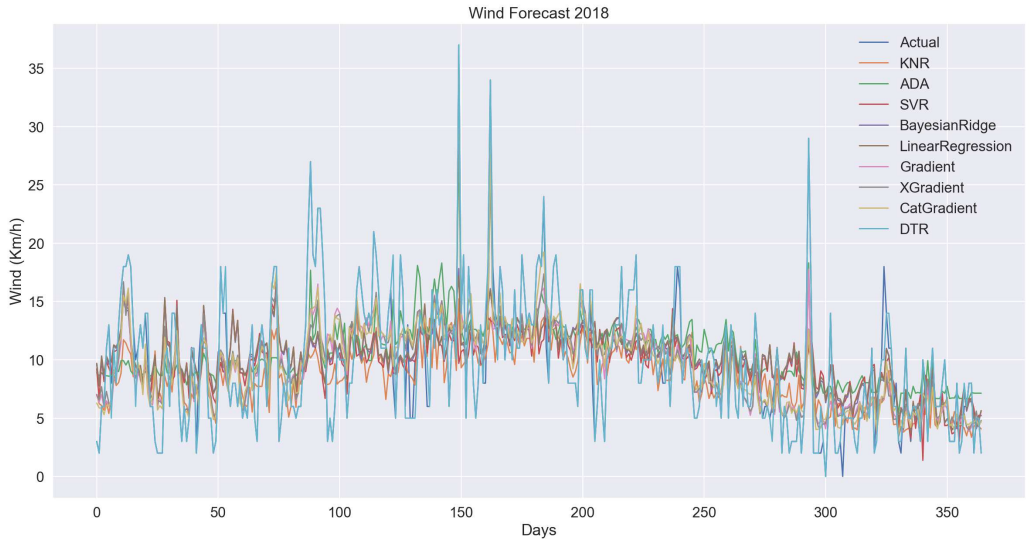

Fig. 19. Forecasting of wind speed for each month of the year for all algorithms compared to original values.

Figure 19 and Figure 20, both figures represent the forecasting performance of our ML-based model where the efficacy of the regression algorithms is noticed. The output shows quite similar performance in case of predicting which proves the reliability of our statistical model. It can be stated from Table 1 that our model learned well from the training data and the degree of learning was examined via observing its performance on testing data. It can be stated from Table 2 that statistical analysis verifies the effectiveness of our weather forecasting model as it could successfully predict the wind speed compared to the actual value.

Table 3 represents the statistical analysis of rainfall for the given dataset of the year 2012-2017. The values of MAE, MSE, and MAPE were measured based on the testing dataset which indicates the learning efficiency of the model. From these measured values we can notice that AdaBoost, Linear Regression

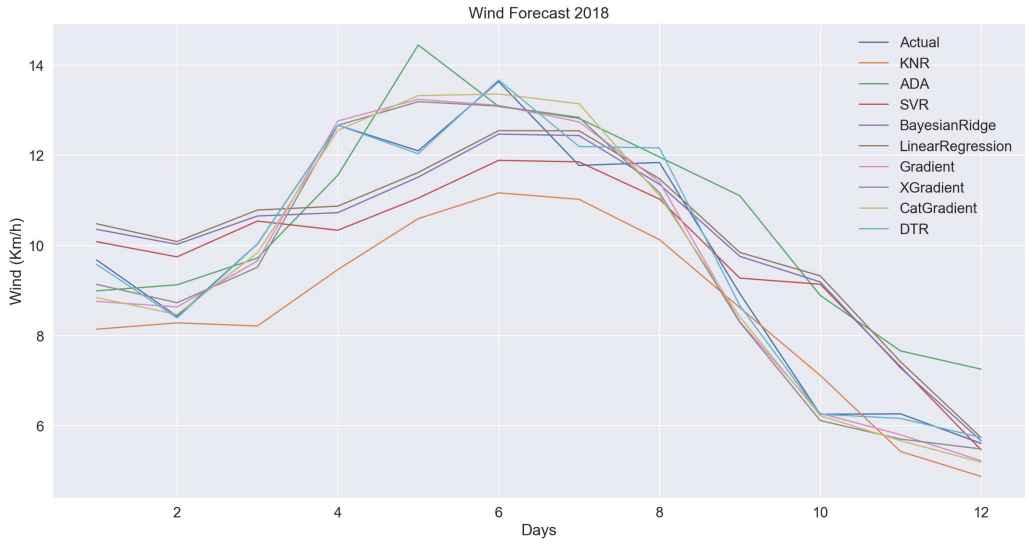

Fig. 20. Forecasting of wind speed for each day of the year for all algorithms compared to original values.

Table 3
Statistical analysis for rainfall for the 2012-2017 dataset.

Algorithms	MAE	MSE	MAPE
KNN	81.42	17019.17	39.72%
XGBoost	73.20	13353.92	35.71%
GB	72.67	13208.35	35.46%
DTR	95.78	33014.21	46.73%
AdaBoost	132.29	28352.98	64.67%
SVR	71.88	15054.07	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%

and Bayesian Ridge showed much more error compared to other algorithms in case of predicting rainfall. Point to be noted that AdaBoost showed the most percentage of error which was approximately 64.67% obtained from MAPE. Initially, the algorithms were trained with train data and later tested using testing data. Amongst all these algorithms CatBoost showed comparatively better performance in case of showing error percentage whose MAPE value was 32.93% and apart from this, rest of the algorithms showed moderate and quite satisfactory performance whose MAPE values were 39.72%, 35.71%, 35.46%, 46.73% and 35.07% respectively for KNN, XGBoost, GB, DTR, and SVR.

Figure 21 depicts the performance comparisons between training and testing data for each of the algorithms. Here the comparisons show that algorithms were properly trained as the blue colored curve determines the actual prediction in training data. Whether the red colored curve expressed the prediction in testing data. From both of the curves, it can be said that the machine-learned properly in order to predict the rainfall which is proved by the output of testing data. The predicted result was almost equal to the actual result of the training data.

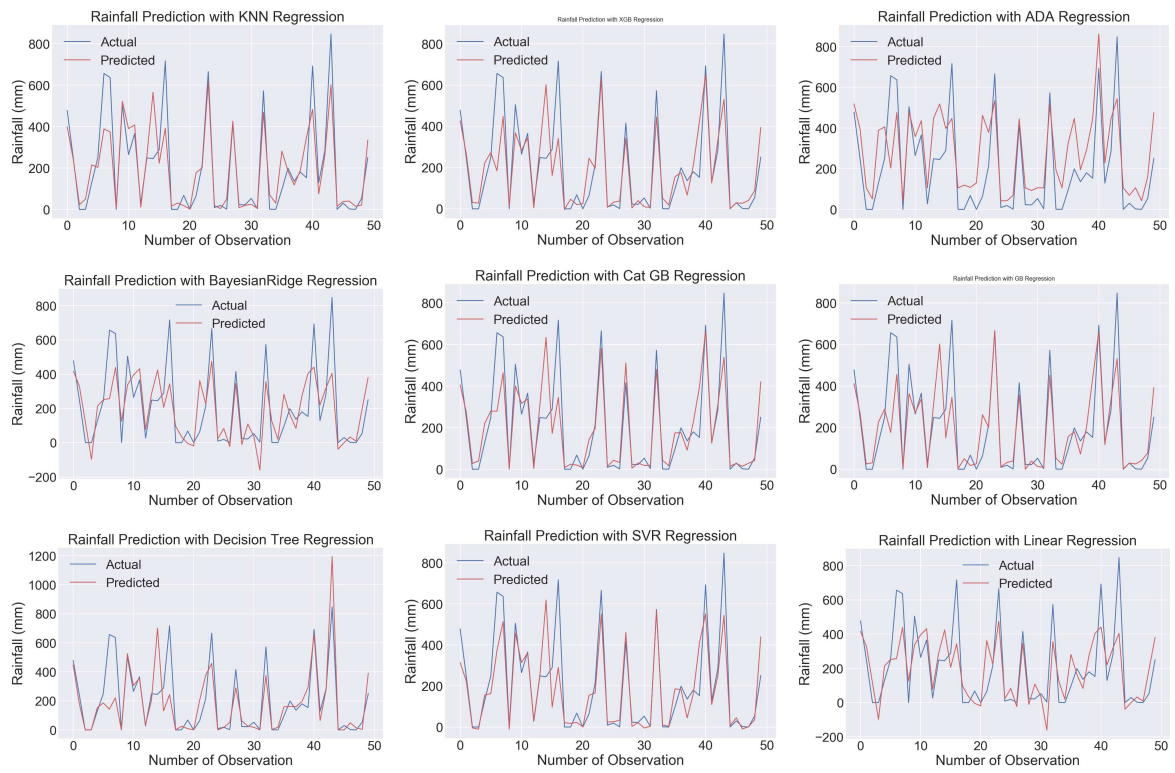

Fig. 21. Performance comparison for randomly split training data and testing data for fifty observation for rainfall prediction using nine ML algorithms(KNN, BR, LR, AdaBoost, GB, XGBoost, CatBoost, DTR, SVR).

Table 4
Statistical analysis for rainfall for the 2018 dataset.

Algorithms	MAE	MSE	MAPE
KNN	87.33	23173.56	46.13%
XGBoost	81.52	19049.68	43.07%
GB	82.67	19480.34	43.67%
DTR	5.17	944.20	2.73%
AdaBoost	159.99	44107.76	84.54%
SVR	86.91	34323.65	45.91%
Linear Regression	121.73	36339.33	64.31%
Bayesian Ridge	121.67	36335.82	64.28%
CatBoost	69.90	14575.92	36.93%

Table 4 is the statistical analysis of the rainfall in case of forecasting for 2018. From the values of MAE, MSE, and MAPE we can notice that DTR performs exceedingly well in forecasting having MAPE of just 2.73% like it did outperform other algorithms in forecasting wind speed. So it proves again that DTR plays a vital role in forecasting due to its structure and algorithm's tree construction. Like the training and testing data AdaBoost showed the least effective performance having the most error percentage of MAPE value which was approximately 84.54%. This was a huge margin considering the

scale it was utilizing in forecasting. Linear Regression and Bayesian Ridge also showed some inconsistent performance having MAPE value of 64.31% and 64.28% which is quite unacceptable observing the performance of other algorithms. Apart from that other algorithms showed moderate performance and their MAPE values were 36.93%, 45.91%, 43.67%, 43.07%, 46.13% respectively for CatBoost, SVR, GB, XGBoost, and KNN.

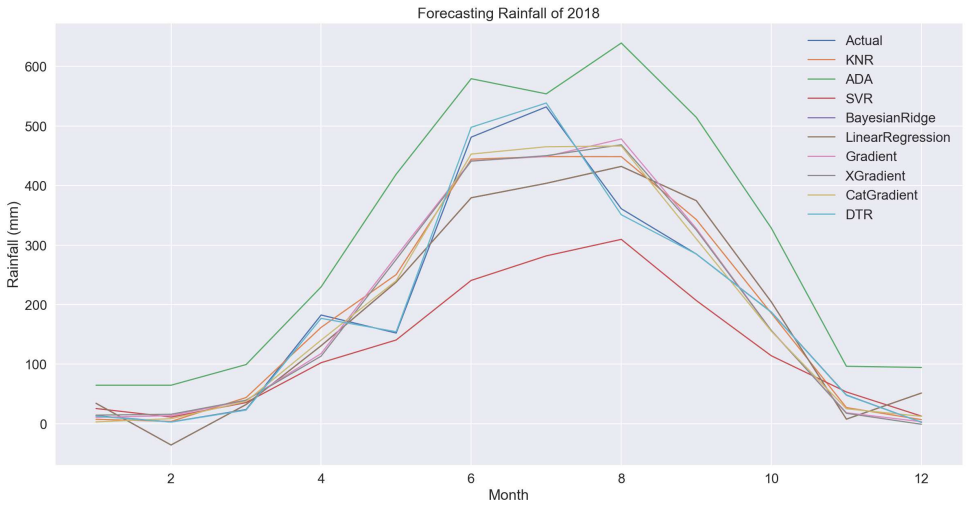

Fig. 22. Forecasting of rainfall for each month of the year for all algorithms compared to original values.

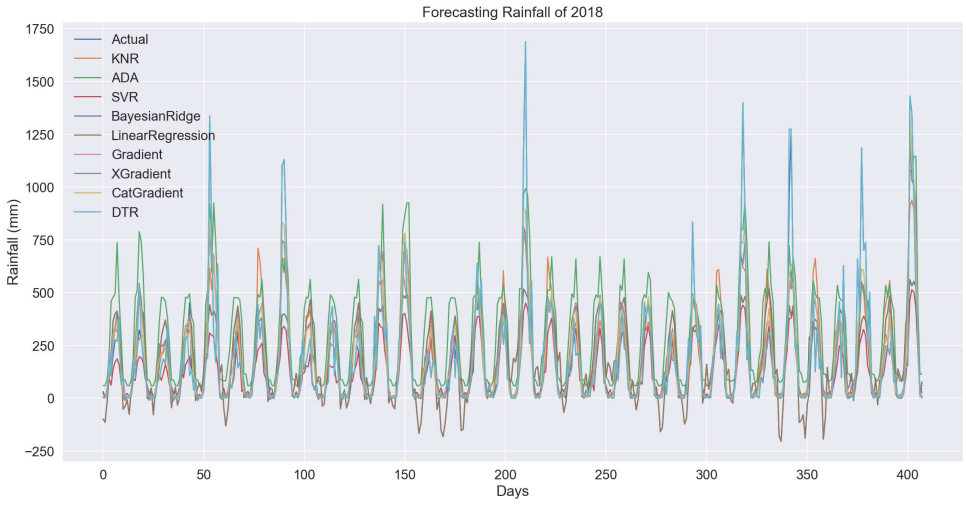

Fig. 23. Forecasting of rainfall for each day of the year for all algorithms compared to original values.

Figure 22 and Figure 23 illustrates the rainfall forecasting performance of our regression-based ML-model. Here Figure 22 shows the monthly forecasting of rainfall for 2018 based on their previous experience from training and testing data. It clearly showed that AdaBoost did not show coherent performance along with other algorithms where SVR also lagged behind others in case of predicting the exact

Table 5
Statistical analysis for humidity for the 2012-2017 dataset.

Algorithms	MAE	MSE	MAPE
KNN	1.29	2.83	1.71%
XGBoost	0.013	0.0015	0.0178%
GB	0.0015	1.92	0.0201%
DTR	0.009	0.009	0.012%
AdaBoost	0.704	0.84	0.93%
SVR	0.031	0.0015	0.041%
Linear Regression	2.18	6.37	2.866×10^{-14}%
Bayesian Ridge	3.21	1.85	4.255×10^{-12}%
CatBoost	1.24	2.76	1.64%

value. In the case of Figure 23, we can notice the daily basis annual rainfall forecasting by our model. It illustrates the combined performance of each algorithm in case of forecasting rainfall on a daily basis of the year. Both figures demonstrate the forecasting performance of our ML-based model where the efficiency of the regression algorithms is visibly noticed. The output shows quite similar performance in case of predicting which proves the stability of our statistical model.

From Table 5, we obtain some impressive results for the statistical analysis for testing data after being trained up. All the algorithms showed almost a hundred percent accuracy in predicting in testing data ignoring the negligible amount of error percentage. The MAE, MSE and MAPE values were too small that we can come to a conclusion that the algorithms learned fully based on the training data and implemented that acquired knowledge successfully on the testing data. Statistically, this statement is proved by the values obtained in Table 5.

It can be demonstrated in Figure 24 that we hardly can separate the curves for prediction in training data and testing data. Most of the cases the actual and predicted lines were overlapped which signifies that the algorithms absorbed the knowledge from the training data properly and implanted it in case of testing data. The obtained statistical values in Table 5 is a mirror reflection of what we achieved in 24. Table 6 can be developed to observe the statistical analysis of forecasting performance in case of the dataset of 2018. Here it is noticed that Table 6 almost resembles Table 5. The values of MAE, MSE, and MAPE were too much negligible in case of forecasting. Significantly, DTR again topped all of the algorithms in this close contest of forecasting. DTR completely had zero errors which made it impressive to count as a perfect regression algorithm to be used in humidity forecasting.

From Figure 25 and Figure 26 we can obtain the monthly and daily humidity forecasting of the year. These two figures significantly reflect what has been achieved in Figure 24, Table 5 and Table 6. The error percentages are too much negligible that we can rarely distinguish the curves of different algorithms. But it carries a huge significance in case of showing the worth of our model. It can be stated that regression-based algorithms hardly put any error in humidity forecasting. We can obtain almost complete accuracy in these algorithms for humidity forecasting which is verified by both of these figures.

Table 7 and Table 8 establishes a verity that regression algorithms show optimum output in case of predicting high temperature. Here Table 7 sets the standard for an initial dataset for testing data based on training data while Table 8 showed performance for the second dataset for forecasting. In both cases, the values of MAE, MSE, and MAPE were very much negligible which proved the authenticity of this model in high temperature predicting. In a nutshell, it can be stated that regression-based ML algorithms provide almost errorless prediction.

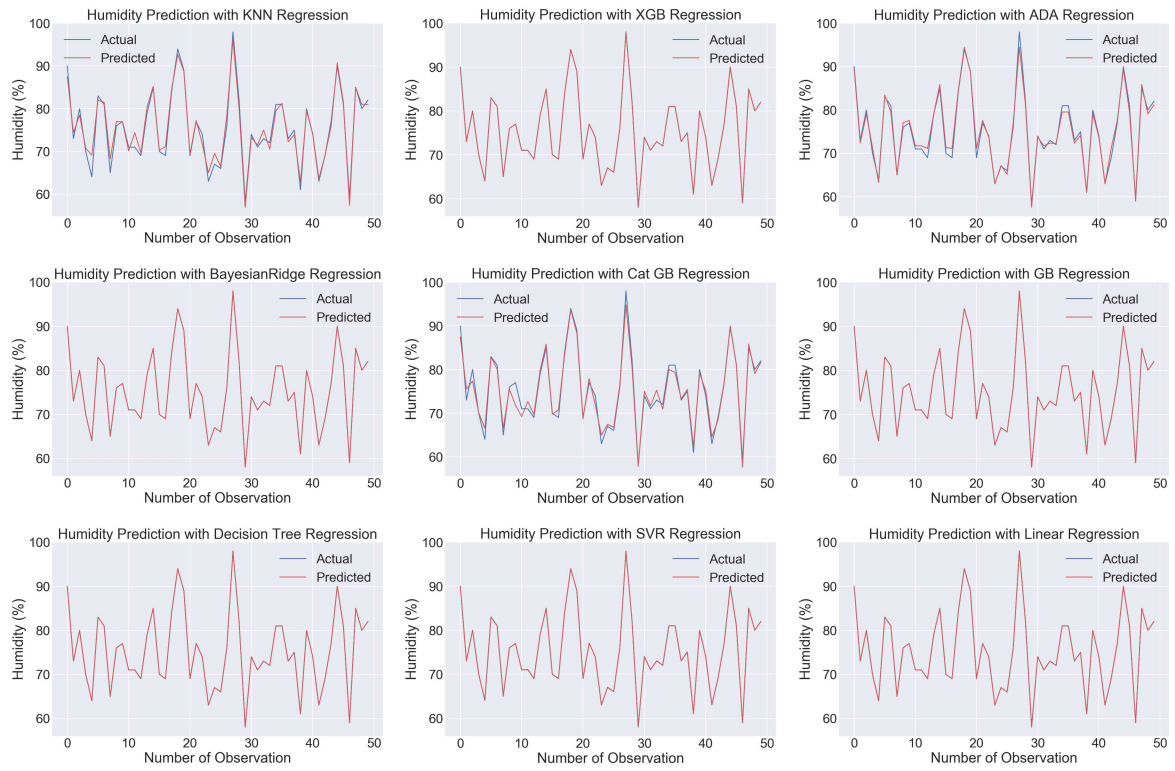

Fig. 24. Performance comparison for randomly split training data and testing data for fifty observation for humidity prediction using nine ML algorithms(KNN, BR, LR, AdaBoost, GB, XGBoost, CatBoost, DTR, SVR).

Table 6
Statistical analysis for humidity for the 2018 dataset.

Algorithms	MAE	MSE	MAPE
KNN	1.82	2.29	1.59%
XGBoost	0.013	0.0011	0.0176%
GB	0.0021	3.26	0.00278%
DTR	0	0	0%
AdaBoost	0.93	1.39	1.25%
SVR	0.0293	0.0013	0.04%
Linear Regression	1.67	4.47	2.25×10^{-14}%
Bayesian Ridge	3.28	1.67	4.42×10^{-12}%
CatBoost	1.10	2.06	1.49%

Figure 27 depicts the similar scenario that we encountered previously for humidity prediction. As the error percentage was minimal for all the algorithms, so the actual and predicted curves overlapped mostly. Due to less error percentage in MAPE values and moreover the MAE and MSE values were also significantly less than conventional methods which promise an opportunistic model for high-temperature forecasting. All the algorithms we used in our model showed sheer corrosiveness which bolstered the claim of a most authentic high-temperature model for Bangladesh. Similarly, we can obtain the monthly

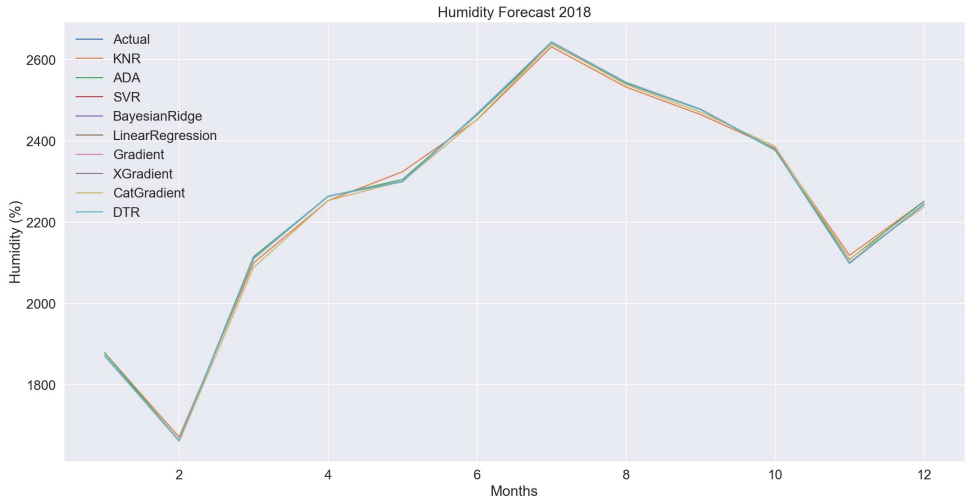

Fig. 25. Forecasting of humidity for each month of the year for all algorithms compared to original values.

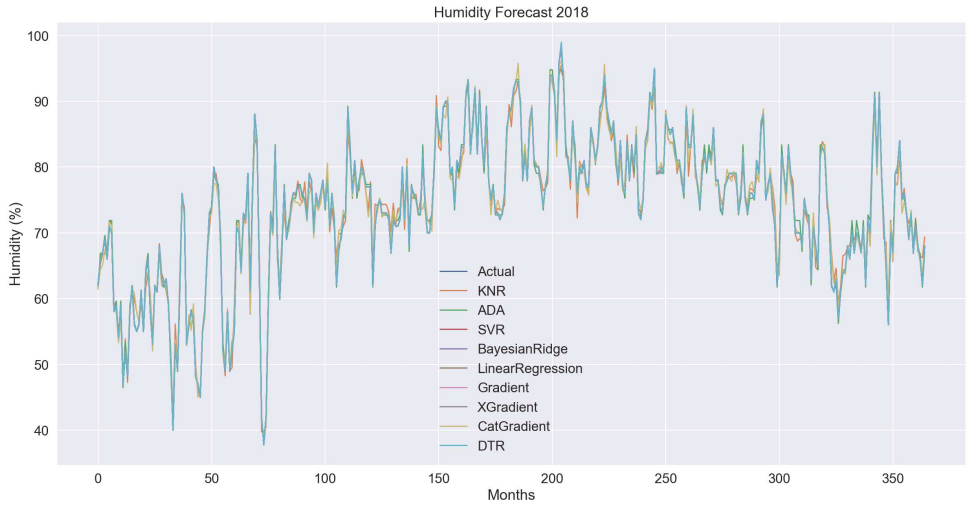

Fig. 26. Forecasting of humidity for each day of the year for all algorithms compared to original values.

Table 7
Statistical analysis for high temperature for the 2012-2017 dataset.

Algorithms	MAE	MSE	MAPE
KNN	0.79	0.98	2.65%
XGBoost	0.0015	4.82	0.0051%
GB	0.00016	5.16	5.23×10^{-4}%
DTR	0	0	0%
AdaBoost	0.24	0.11	0.80%
SVR	0.21	0.11	0.67%
Linear Regression	2.08	6.07	7.03×10^{-14}%
Bayesian Ridge	1.16	2.03	3.91×10^{-11}%
CatBoost	0.43	0.30	1.44%

Table 8
Statistical analysis for high temperature for the 2018 dataset.

Algorithms	MAE	MSE	MAPE
KNN	0.78	1.11	2.54%
XGBoost	0.0016	5.51	5.29×10^{-3}%
GB	0.00016	5.38	5.05×10^{-4}%
DTR	0	0	0%
AdaBoost	0.27	0.13	0.87%
SVR	0.032	0.0016	0.10%
Linear Regression	1.96×10^{-14}	5.67×10^{-28}	6.36×10^{-14}%
Bayesian Ridge	1.18×10^{-11}	2.14×10^{-22}	3.85×10^{-11}%
CatBoost	0.41	0.27	1.30%

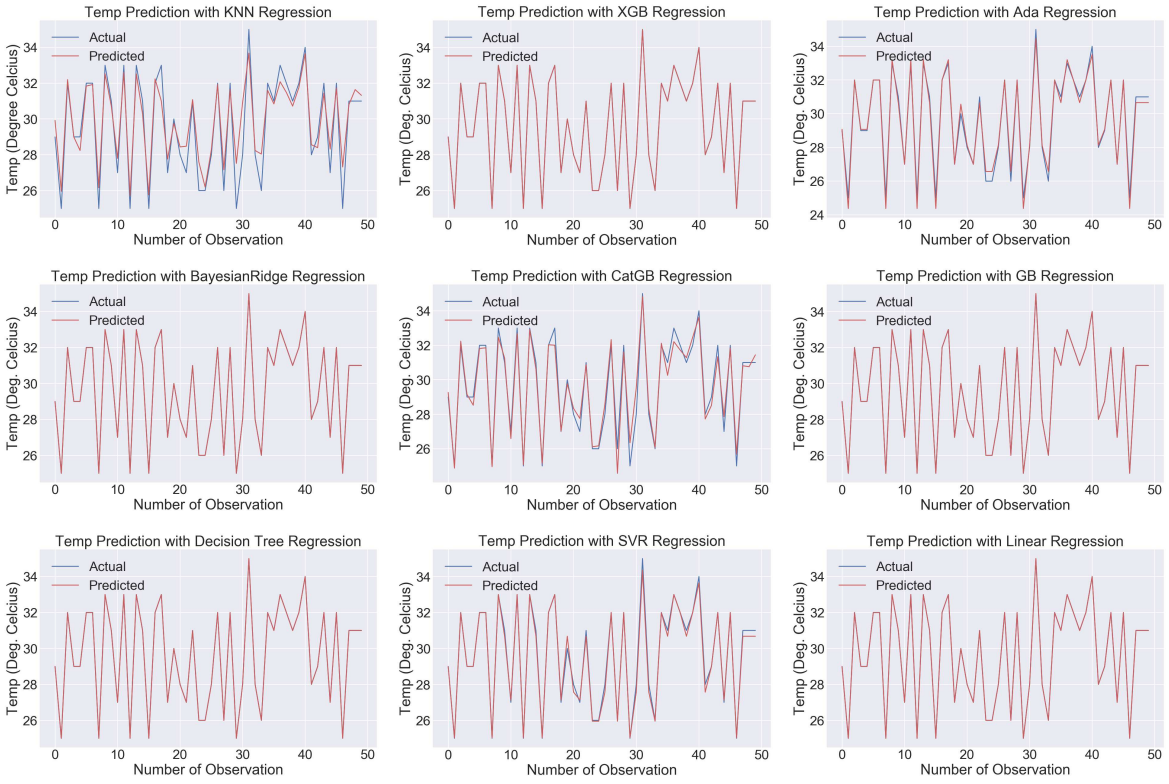

Fig. 27. Performance comparison for randomly split training data and testing data for fifty observation for high temperature prediction using nine ML algorithms(KNN, BR, LR, AdaBoost, GB, XGBoost, CatBoost, DTR, SVR).

and daily high-temperature forecasting observations from our model and compare it with the actual values which are preserved in the dataset of 2018. It can be noticed from Figure 28 and Figure 29 that the outputs of our used algorithms don't differ that much from the actual value which ensures the forecasting accuracy of our model. Figure 28 represents the monthly high-temperature forecast of 2018 whether Figure 29 illustrates the high-temperature forecast for each day of 2018. From Figure 27,

Table 9
Statistical analysis for low temperature for the 2012-2017 dataset.

Algorithms	MAE	MSE	MAPE
KNN	0.91	1.37	4.04%
XGBoost	0.00063	4.74	2.85×10^{-3}%
GB	0.00021	7.82	8.83×10^{-4}%
DTR	0	0	0%
AdaBoost	0.44	0.21	1.97%
SVR	0.033	0.0018	0.15%
Linear Regression	2.41×10^{-14}	7.84×10^{-28}	1.08×10^{-11}%
Bayesian Ridge	1.16×10^{-11}	2.08×10^{-22}	5.23×10^{-11}%
CatBoost	0.42	0.32	1.87%

Table 7 and Table 8, 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 showed better performance compared to others.

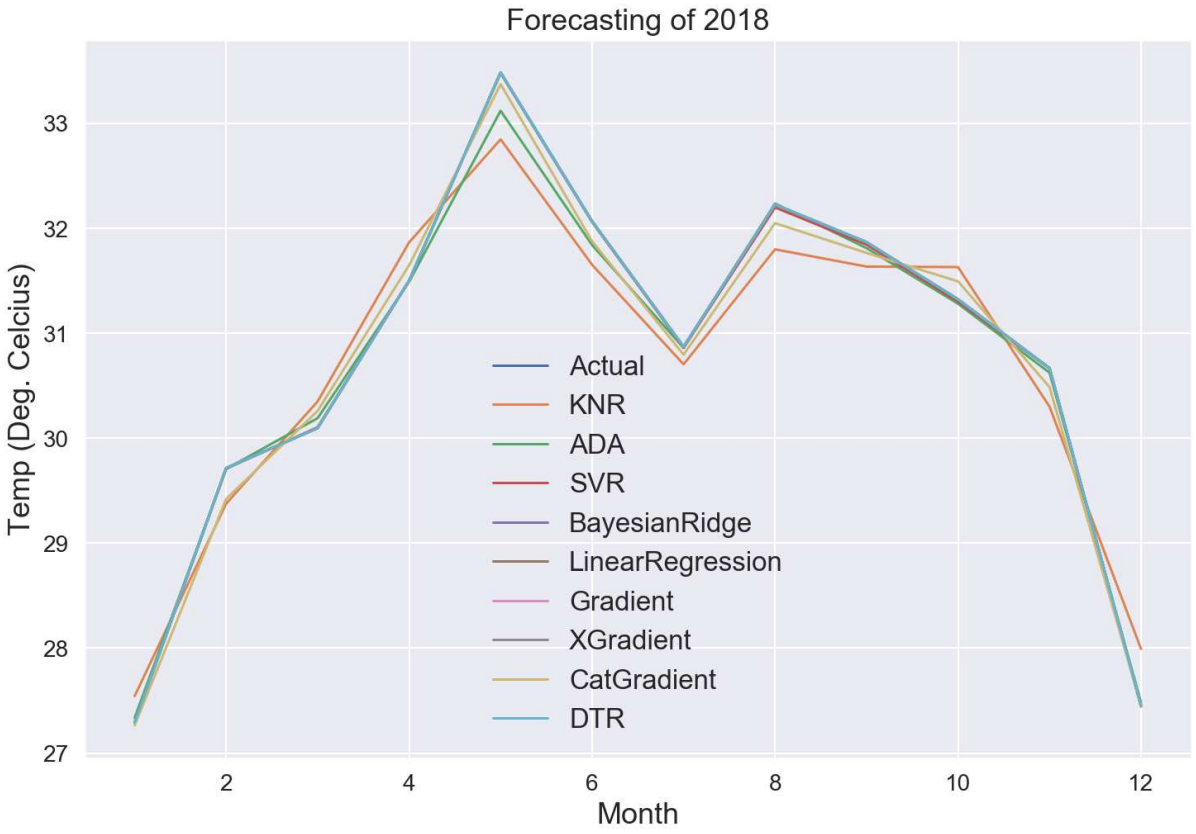

Fig. 28. Forecasting of high temperature for each month of the year for all algorithms compared to original values.

From the above-mentioned Table 9 and Table 10, we can monitor the statistical analysis and obtained values of MAE, MSE, and MAPE. Like the high-temperature prediction and forecasting here we also

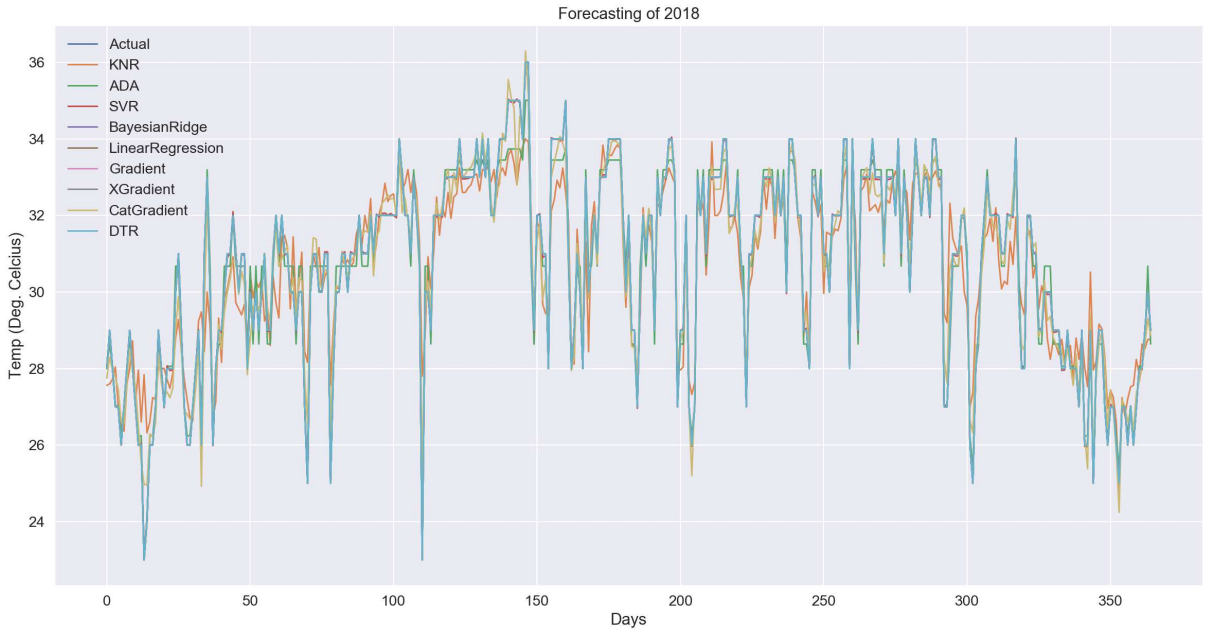

Fig. 29. Forecasting of high temperature for each day of the year for all algorithms compared to original values.

Table 10
Statistical analysis for low temperature for the 2018 dataset.

Algorithms	MAE	MSE	MAPE
KNN	0.81	1.21	3.56%
XGBoost	0.00063	4.57	2.78×10^{-3}%
GB	0.00016	5.35	7.10×10^{-4}%
DTR	0	0	0%
AdaBoost	0.37	0.16	1.62%
SVR	0.04	0.0023	0.18%
Linear Regression	2×10^{-14}	6.41×10^{-28}	8.81×10^{-14}%
Bayesian Ridge	1.17×10^{-11}	2.10×10^{-22}	5.14×10^{-11}%
CatBoost	0.38	0.26	1.65%

found that error percentage is minimal compared to other conventional methods and DTR shows the best performance amongst all the algorithms. Once again DTR provided zero percentage of 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 percentage of error. Comparing both of these tables we found that KNN provided the most percentage of error in low-temperature forecasting. Its MAPE values were 4.04% and 3.56% respectively for the initial dataset and 2018's forecasting dataset.

Figure 30 portrays the comparable situation that we experienced beforehand for high-temperature expectation. As the error rate was negligible for every one of the calculations, so the real and anticipated bends covered for the most part. Because of less blunder rate in MAPE values and in addition, the MAE and MSE values were additionally altogether not exactly traditional strategies which guarantee a shrewd model for low-temperature anticipating. Every one of the calculations we utilized in our model indi-

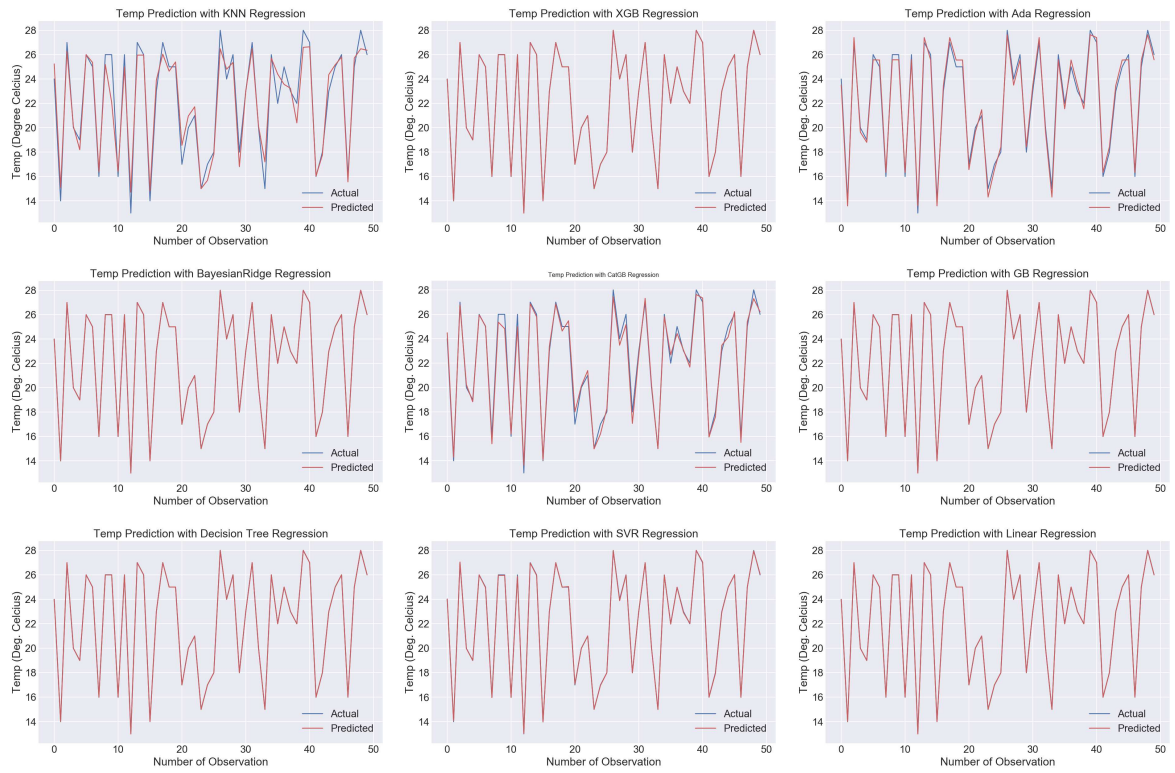

Fig. 30. Performance comparison for randomly split training data and testing data for fifty observation for low temperature prediction using nine ML algorithms(KNN, BR, LR, AdaBoost, GB, XGBoost, CatBoost, DTR, SVR).

cated sheer destructiveness which supported the case of most credible low-temperature demonstrate for Bangladesh. Likewise, we can acquire the month to month and day by day low-temperature determining perceptions from our model and contrast it and the genuine qualities which are saved in the dataset of 2018. It very well may be seen from Figure 31 and Figure 32 that the yields of our utilized calculations don't contrast that much from the real esteem which guarantees the gauging exactness of our model. Figure 31 speaks to the month to month low-temperature conjecture of 2018 whether Figure 32 outlines the low-temperature estimate of every day of 2018. From Figure 30, Table 9 and Table 10, it tends to be expressed that DTR demonstrated the best execution in demonstrating errorless low-temperature anticipating. Aside from DTR, Linear Regression and Bayesian Ridge indicated better execution contrasted with others. KNN showed comparatively more errors than other algorithms in this analysis. Overall, it is proven with exclusive evidence that ML-based algorithms show better performance in terms of predicting the wind-speed, rainfall, humidity, high temperature, and low temperature.

4. Conclusion and Future Works

In this paper, it is tried to introduce the regression algorithms in order to handle the uphill task weather events forecasting including the weather events like rainfall, wind speed, humidity, high and low temperature using the provided climatic information. Our findings suggest that regression-based ML-algorithms

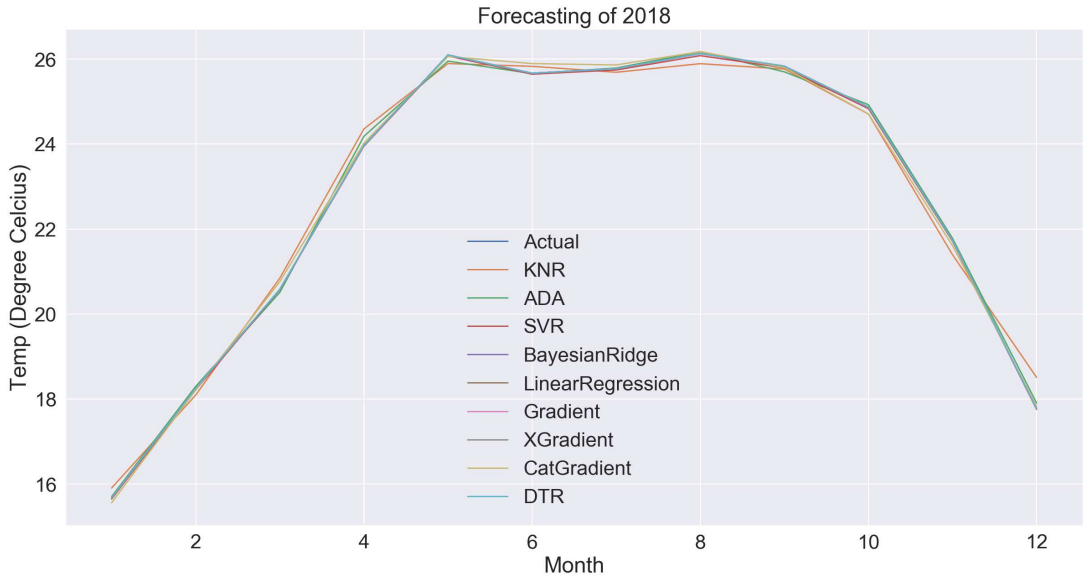

Fig. 31. Forecasting of low temperature for each month of the year for all algorithms compared to original values.

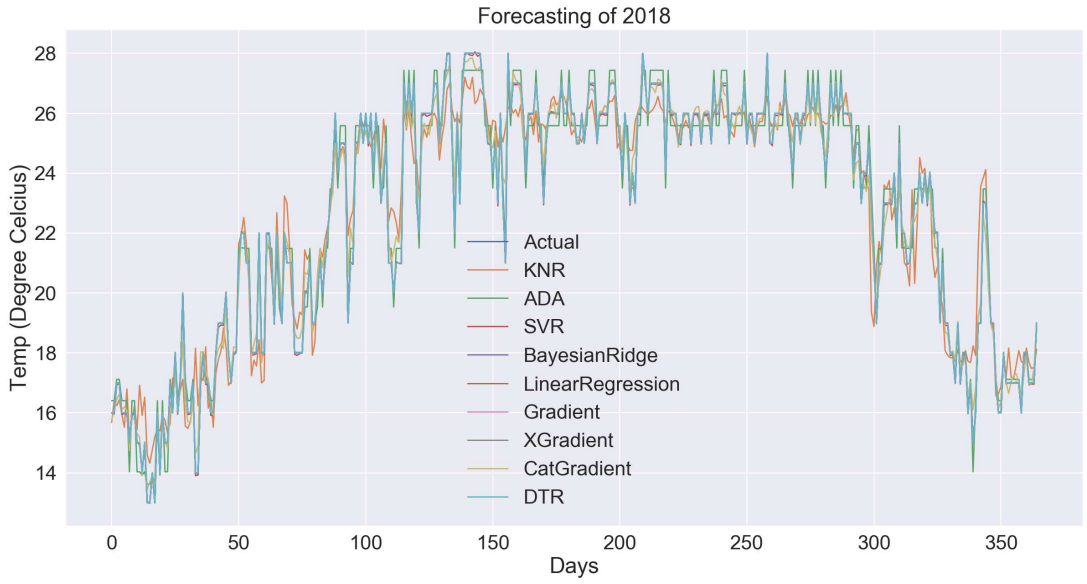

Fig. 32. Forecasting of low temperature for each day of the year for all algorithms compared to original values.

are capable of predicting weather events with a narrow margin of error rate utilizing the less number of weather parameters. Our proposed model was trained to forecast those above mentioned specific weather parameters monthly and daily basis of the year. Initially, the daily, monthly and annual amount of those weather events were plotted using the recorded data from 2012 to 2017. Later they were trained for all of those algorithms using the training dataset and were tested using the testing dataset for each of the specific weather events. Finally, the forecasting was monitored in case of 2018 for the monthly and daily

basis of the year. The performance was evaluated for four of the mentioned weather parameters and compared for all of those algorithms. The comparison of the proposed model with many other state-of-the-art strategies, and the approval of the proposed weather forecasting model on more mind-boggling and uneven weather dataset with several climate conditions.

For weather forecasting, many techniques and types are available. There is a lot of methods to predict the weather events e.g. rainfall, humidity, wind speed, high & low temperature. Some are often used methods (Linear Regression, Bayesian Ridge, and DTR) in case of weather forecasting. Other more frequently used algorithms (SVR, KNN) and rarely used boosting algorithms (GB, XGBoost, AdaBoost, CatBoost) were also applied to weather data to predict the events successfully. In some cases, one was proved to be the best and in others, it was the reverse. As a conclusion, it can be said that DTR and CatBoost methods were almost equivalent in terms of quality of prediction in certain fluctuation condition, but the adaptability of DTR as widespread nonlinear assumption makes them more idea than CatBoost. Generally, the accuracy of these methods is hugely dependent on the quality of the training dataset. Actually, considering all the aspects, these methods contain some comparable statistical error. The execution of the strategies may have more to do with the mistakes revealed in the writing than the techniques themselves. For instance, when the auto-correlation of the error is decreased white noise for the similar data sources, SVR or regression trees (DTR) perform in all respects likewise, with no factual contrasts between them. The other fact is that ensemble methodologies are always better than simple predictors. In the present paper, an effort has been put to build an ML-based weather forecasting model for Bangladesh based on a set of regression algorithms that attempt to model high-level abstraction in data by using the model architecture, with complicated structures or otherwise.

Although, this research area is considered to be a neophyte and enough experiences are not available. But in the near future, these sorts of models might completely dominate over the conventional methods which are already in use. Consequently, forecasts obtained via several methods can be measured to satiate the various requirements. In this paper, it was tried to forecast rainfall, humidity, wind speed, high and low temperature but in the long run, we can add some other complicated features to the tail. Some complex features like dew point computing, rain-fog, thunderstorm, rain-thunderstorm, tornado, rain-tornado, rain-thunderstorm-tornado, fog-rain-thunderstorm, etc. can be extended to the list of predicting. The question might arise that how they would be assembled together. The appropriate response is obviously not paltry since the different coming about conjectures show contrasts on numerous focuses. Besides, some of them will be related to certainty interim which ought to likewise be consolidated.

References

[1] M.N. Islam and H. Uyeda, Use of TRMM in determining the climatic characteristics of rainfall over Bangladesh, *Remote Sensing of Environment* **108**(3) (2007), 264–276.

[2] M.M.H. Khan, I. Bryceson, K.N. Kolivras, F. Faruque, M.M. Rahman and U. Haque, Natural disasters and land-use/land-cover change in the southwest coastal areas of Bangladesh, *Regional Environmental Change* **15**(2) (2015), 241–250.

[3] R. Roy and N.W. Chan, An assessment of agricultural sustainability indicators in Bangladesh: review and synthesis, *The Environmentalist* **32**(1) (2012), 99–110.

[4] M.A. Rahman, L. Yunsheng and N. Sultana, Analysis and prediction of rainfall trends over Bangladesh using Mann–Kendall, Spearman's rho tests and ARIMA model, *Meteorology and Atmospheric Physics* **129**(4) (2017), 409–424.

[5] M.C. Serreze and R.G. Barry, Processes and impacts of Arctic amplification: A research synthesis, *Global and planetary change* **77**(1–2) (2011), 85–96.

[6] R. Paull, Effect of temperature and relative humidity on fresh commodity quality, *Postharvest biology and technology* **15**(3) (1999), 263–277.

[7] S. Al-Yahyai, Y. Charabi and A. Gastli, Review of the use of numerical weather prediction (NWP) models for wind energy assessment, *Renewable and Sustainable Energy Reviews* **14**(9) (2010), 3192–3198.
[8] M. Fernández-Fernández, M. Gallego, F. Domínguez-Castro, R. Trigo, J. García, J. Vaquero, J.M. González and J.C. Durán, The climate in Zafra from 1750 to 1840: history and description of weather observations, *Climatic change* **126**(1–2) (2014), 107–118.
[9] A.B. Smith and R.W. Katz, US billion-dollar weather and climate disasters: data sources, trends, accuracy and biases, *Natural hazards* **67**(2) (2013), 387–410.
[10] O. Coddington, J. Lean, P. Pilewskie, M. Snow and D. Lindholm, A solar irradiance climate data record, *Bulletin of the American Meteorological Society* **97**(7) (2016), 1265–1282.
[11] F. Olaiya and A.B. Adeyemo, Application of data mining techniques in weather prediction and climate change studies, *International Journal of Information Engineering and Electronic Business* **4**(1) (2012), 51.
[12] C. Rudin and K.L. Wagstaff, Machine learning for science and society, Springer, 2014.
[13] L. Delle Monache, F.A. Eckel, D.L. Rife, B. Nagarajan and K. Searight, Probabilistic weather prediction with an analog ensemble, *Monthly Weather Review* **141**(10) (2013), 3498–3516.
[14] S. Johnson, F. Fielding, G. Hamilton and K. Mengersen, An integrated Bayesian network approach to Lyngbya majuscula bloom initiation, *Marine Environmental Research* **69**(1) (2010), 27–37.
[15] S.J. Moe, S. Haande and R.-M. Couture, Climate change, cyanobacteria blooms and ecological status of lakes: a Bayesian network approach, *Ecological modelling* **337** (2016), 330–347.
[16] W.M. Dlamini, A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland, *Environmental Modelling & Software* **25**(2) (2010), 199–208.
[17] A.-Z.S.B. Habib, S. Mallick, A.S. Ahmed, S.S. Alam and A.S. Ahmad, Performance Appraisal of Spectrum Sensing in Cognitive Radio Network, in: *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, IEEE, 2018, pp. 162–167.
[18] C.H. Lima and U. Lall, Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow, *Journal of Hydrology* **383**(3–4) (2010), 307–318.
[19] J. Wu and E. Chen, A novel nonparametric regression ensemble for rainfall forecasting using particle swarm optimization technique coupled with artificial neural network, in: *International Symposium on Neural Networks*, Springer, 2009, pp. 49–58.
[20] B. Pradhan and A.M. Youssef, Manifestation of remote sensing data and GIS on landslide hazard analysis using spatial-based statistical models, *Arabian Journal of Geosciences* **3**(3) (2010), 319–326.
[21] X. Li, H. Xie, R. Wang, Y. Cai, J. Cao, F. Wang, H. Min and X. Deng, Empirical analysis: stock market prediction via extreme learning machine, *Neural Computing and Applications* **27**(1) (2016), 67–78.
[22] R.F. Chevalier, G. Hoogenboom, R.W. McClendon and J.A. Paz, Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks, *Neural Computing and Applications* **20**(1) (2011), 151–159.
[23] K. Abhishek, M. Singh, S. Ghosh and A. Anand, Weather forecasting model using artificial neural network, *Procedia Technology* **4** (2012), 311–318.
[24] D.I. Jeong, A. St-Hilaire, T.B. Ouarda and P. Gachon, Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator, *Climatic Change* **114**(3–4) (2012), 567–591.
[25] J.M.L. Sloughter, A.E. Raftery, T. Gneiting and C. Fraley, Probabilistic quantitative precipitation forecasting using Bayesian model averaging, *Monthly Weather Review* **135**(9) (2007), 3209–3220.
[26] M. Taillardat, O. Mestre, M. Zamo and P. Naveau, Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics, *Monthly Weather Review* **144**(6) (2016), 2375–2393.
[27] C. Persson, P. Bacher, T. Shiga and H. Madsen, Multi-site solar power forecasting using gradient boosted regression trees, *Solar Energy* **150** (2017), 423–436.
[28] P. Karvelis, S. Kolios, G. Georgoulas and C. Stylios, Ensemble learning for forecasting main meteorological parameters, in: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2017, pp. 3711–3714.
[29] S.B. Taieb and R.J. Hyndman, A gradient boosting approach to the Kaggle load forecasting competition, *International journal of forecasting* **30**(2) (2014), 382–394.
[30] M.S. Uddin and T. Shioyama, Bipolarity and projective invariant-based zebra-crossing detection for the visually impaired, in: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, IEEE, 2005, pp. 22–22.
[31] T.-w. Kim, H. Ahn, G. Chung and C. Yoo, Stochastic multi-site generation of daily rainfall occurrence in south Florida, *Stochastic Environmental Research and Risk Assessment* **22**(6) (2008), 705–717.
[32] R.F. Chevalier, G. Hoogenboom, R.W. McClendon and J.A. Paz, Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks, *Neural Computing and Applications* **20**(1) (2011), 151–159.

[33] E.J. Gill, E.B. Singh and E.S. Singh, Training back propagation neural networks with genetic algorithm for weather forecasting, in: *IEEE 8th International Symposium on Intelligent Systems and Informatics*, IEEE, 2010, pp. 465–469.
[34] F. Olaiya and A.B. Adeyemo, Application of data mining techniques in weather prediction and climate change studies, *International Journal of Information Engineering and Electronic Business* **4**(1) (2012), 51.
[35] D. Luminto, Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate, in: *2017 International Conference, Concepts, Theory, and Applications (ICAICTA)*, 2017.
[36] N. Sharma, P. Sharma, D. Irwin and P. Shenoy, Predicting solar generation from weather forecasts using machine learning, in: *2011 IEEE international conference on smart grid communications (SmartGridComm)*, IEEE, 2011, pp. 528–533.
[37] S.A. Nishe, T.A. Tahrin, N. Kamal, M.S. Hoque and K.T. Hasan, Micro-level meteorological data sourcing for accurate weather prediction, in: *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, IEEE, 2017, pp. 353–356.
[38] Y. Zaman, Machine Learning Model on Rainfall-A Predicted Approach for Bangladesh, PhD thesis, United International University, 2018.
[39] N. Hasan, M.T. Uddin and N.K. Chowdhury, Automated weather event analysis with machine learning, in: *2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, IEEE, 2016, pp. 1–5.
[40] K. Mohammadi, S. Shamshirband, S. Motamedi, D. Petković, R. Hashim and M. Gocic, Extreme learning machine based prediction of daily dew point temperature, *Computers and Electronics in Agriculture* **117** (2015), 214–225.
[41] N. Hasan, N.C. Nath and R.I. Rasel, A support vector regression model for forecasting rainfall, in: *2015 2nd International Conference on Electrical Information and Communication Technologies (EICT)*, IEEE, 2015, pp. 554–559.
[42] V.M. Krasnopolsky and M.S. Fox-Rabinovitz, Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction, *Neural Networks* **19**(2) (2006), 122–134.
[43] A. Ata, M. Khan, S. Abbas, G. Ahmad and A. Fatima, MODELLING SMART ROAD TRAFFIC CONGESTION CONTROL SYSTEM USING MACHINE LEARNING TECHNIQUES, *Neural Network World* **29**(2) (2019), 99–110.
[44] A. Atta, S. Abbas, M.A. Khan, G. Ahmed and U. Farooq, An adaptive approach: Smart traffic congestion control system, *Journal of King Saud University-Computer and Information Sciences* (2018).
[45] A. Saeed, M. Yousif, A. Fatima, S. Abbas, M. Adnan Khan, L. Anum and A. Akram, An Optimal Utilization of Cloud Resources using Adaptive Back Propagation Neural Network and Multi-Level Priority Queue Scheduling, *The ISC International Journal of Information Security* **11**(3) (2019), 145–151.
[46] A. Fatima, M.A. Sagheer Abbas, M.A. Khan and M.S. Khan, Optimization of Governance Factors for Smart City Through Hierarchical Mamdani Type-1 Fuzzy Expert System Empowered with Intelligent Data Ingestion Techniques (2019).
[47] C.J. Willmott and K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate research* **30**(1) (2005), 79–82.
[48] T. Chai and R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature, *Geoscientific model development* **7**(3) (2014), 1247–1250.
[49] Z. Wang and A.C. Bovik, Mean squared error: Love it or leave it? A new look at signal fidelity measures, *IEEE signal processing magazine* **26**(1) (2009), 98–117.
[50] A. De Myttenaere, B. Golden, B. Le Grand and F. Rossi, Mean absolute percentage error for regression models, *Neurocomputing* **192** (2016), 38–48.
[51] M. Akter, M.S. Uddin and A. Haque, Diagnosis and management of diabetes mellitus through a knowledge-based system, in: *13th International Conference on Biomedical Engineering*, Springer, 2009, pp. 1000–1003.
[52] J. Heo and J.Y. Yang, AdaBoost based bankruptcy forecasting of Korean construction companies, *Applied soft computing* **24** (2014), 494–499.
[53] H. Yang, L. Chan and I. King, Support vector machine regression for volatile stock market prediction, in: *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, 2002, pp. 391–396.
[54] M.G. Blum and O. François, Non-linear regression models for Approximate Bayesian Computation, *Statistics and Computing* **20**(1) (2010), 63–73.
[55] A. Goia, C. May and G. Fusai, Functional clustering and linear regression for peak load forecasting, *International Journal of Forecasting* **26**(4) (2010), 700–711.
[56] J. Huang and M. Perry, A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting, *International Journal of Forecasting* **32**(3) (2016), 1081–1086.
[57] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang and Y. Si, A data-driven design for fault detection of wind turbines using random forests and XGboost, *IEEE Access* **6** (2018), 21020–21031.
[58] G. Huang, L. Wu, X. Ma, W. Zhang, J. Fan, X. Yu, W. Zeng and H. Zhou, Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions, *Journal of Hydrology* **574** (2019), 1029–1041.
[59] A. Ghasemzadeh, B.E. Hammit, M.M. Ahmed and R.K. Young, Parametric ordinal logistic regression and non-parametric decision tree approaches for assessing the impact of weather conditions on driver speed selection using naturalistic driving data, *Transportation research record* **2672**(12) (2018), 137–147.

[60] M. Rajeevan, D. Pai, R.A. Kumar and B. Lal, New statistical models for long-range forecasting of southwest monsoon rainfall over India, *Climate Dynamics* **28**(7–8) (2007), 813–828.
[61] A. Mahabub, M.I. Mahmud and M.F. Hossain, A Robust System for Message Filtering Using an Ensemble Machine Learning Supervised Approach, *ICIC Express Letters, Part B: Applications* **10**(9) (2019), 805–812.
[62] J.R. Andrade and R.J. Bessa, Improving renewable energy forecasting with a grid of numerical weather predictions, *IEEE Transactions on Sustainable Energy* **8**(4) (2017), 1571–1580.
[63] J.R. Andrade and R.J. Bessa, Improving renewable energy forecasting with a grid of numerical weather predictions, *IEEE Transactions on Sustainable Energy* **8**(4) (2017), 1571–1580.
[64] E.C. Morgan, M. Lackner, R.M. Vogel and L.G. Baise, Probability distributions for offshore wind speeds, *Energy Conversion and Management* **52**(1) (2011), 15–26.
[65] A.D. Penwarden, Acceptable wind speeds in towns, *Building Science* **8**(3) (1973), 259–267.
[66] H. do Nascimento Camelo, P.S. Lucio, J.B.V.L. Junior, P.C.M. de Carvalho and D.v.G. dos Santos, Innovative hybrid models for forecasting time series applied in wind generation based on the combination of time series models with artificial neural networks, *Energy* **151** (2018), 347–357.
[67] T.G. Huntington, Evidence for intensification of the global water cycle: review and synthesis, *Journal of Hydrology* **319**(1–4) (2006), 83–95.
[68] J. Le Marshall, J. Jung, J. Derber, M. Chahine, R. Treadon, S.J. Lord, M. Goldberg, W. Wolf, H.C. Liu, J. Joiner et al., Improving global analysis and forecasting with AIRS, *Bulletin of the American Meteorological Society* **87**(7) (2006), 891–895.
[69] S. Sherwood, R. Roca, T. Weckwerth and N. Andronova, Tropospheric water vapor, convection, and climate, *Reviews of Geophysics* **48**(2) (2010).
[70] G. Bontempi, S.B. Taieb and Y.-A. Le Borgne, Machine learning strategies for time series forecasting, in: *European business intelligence summer school*, Springer, 2012, pp. 62–77.
[71] R. Carbonneau, K. Laframboise and R. Vahidov, Application of machine learning techniques for supply chain demand forecasting, *European Journal of Operational Research* **184**(3) (2008), 1140–1154.
[72] D.B. Lobell and M.B. Burke, Why are agricultural impacts of climate change so uncertain? The importance of temperature relative to precipitation, *Environmental Research Letters* **3**(3) (2008), 034007.
[73] J. Mailhot, J. Milbrandt, A. Giguère, R. McTaggart-Cowan, A. Erfani, B. Denis, A. Glazer and M. Vallée, An experimental high-resolution forecast system during the Vancouver 2010 Winter Olympic and Paralympic Games, *Pure and Applied Geophysics* **171**(1–2) (2014), 209–229.
[74] M.T. Chahine, T.S. Pagano, H.H. Aumann, R. Atlas, C. Barnet, J. Blaisdell, L. Chen, M. Divakarla, E.J. Fetzer, M. Goldberg et al., AIRS: Improving weather forecasting and providing new data on greenhouse gases, *Bulletin of the American Meteorological Society* **87**(7) (2006), 911–926.
[75] R.K. Lai, C.-Y. Fan, W.-H. Huang and P.-C. Chang, Evolving and clustering fuzzy decision tree for financial time series data forecasting, *Expert Systems with Applications* **36**(2) (2009), 3761–3773.