Discover Environment

Discover

Research

A perceptible stacking ensemble model for air temperature prediction in a tropical climate zone

Tajrian Mollick[1] · **Galib Hashmi**[2] · **Saifur Rahman Sabuj**[1]

Received: 8 August 2023 / Accepted: 18 September 2023
Published online: 28 September 2023

Abstract
Bangladesh is one of the world's most susceptible countries to climate change. Global warming has significantly increased surface temperatures worldwide, including in Bangladesh. According to meteorological observations, the average temperature of the world has risen approximately 1.2 °C to 1.3 °C over the last century. Researchers and decision-makers have recently paid attention into the climate change studies. Climate models are used extensively throughout the nation in studies on global climate change to determine future estimates and uncertainties. This paper outlines a perceptible stacking ensemble learning model to estimate the temperature of a tropical region—Cox's Bazar, Bangladesh. The next day's temperature, maximum temperature, and minimum temperature are estimated based on the daily weather database collected from the weather station of Cox's Bazar for a period of 20 years between 2001 and 2021. Five machine learning (ML) models, namely linear regression (LR), ridge, support vector regression (SVR), random forest (RF), and light gradient boosting machine (LGBM) are selected out of twelve ML models and combined to integrate the outputs of each model to attain the desired predictive performance. Different statistical schemes based on time-lag values play a significant role in the feature engineering stage. Evaluation metrics like mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) are determined to compare the predictive performance of the models. The findings imply that the stacking approach presented in this paper prevails over the standalone models. Specifically, the study reached the highest attainable R^2 values (0.925, 0.736, and 0.965) for forecasting temperature, maximum temperature, and minimum temperature. The statistical test and trend analysis provide additional evidence of the excellent performance of the suggested model.

Keywords Daily temperature · Maximum temperature · Minimum temperature · Machine learning · Stacking ensemble learning

1 Introduction

Climate change is a major environmental threat affecting many countries globally. Food production, water availability, forest biodiversity, and livelihoods highly relate to it. According to projections made by the Intergovernmental Panel on Climate Change (IPCC), the effects of global warming on human society and the environment would vary throughout time and space [1, 2]. When considering the effects of climate change on Earth and its atmosphere, the temperature parameter is often considered to be the most important of all meteorological variables [1]. Over the last 100 years, the average air temperature near Earth's surface has risen by little under 1 °C. Global warming alters

✉ Saifur Rahman Sabuj, s.r.sabuj@ieee.org | [1]Department of Electrical and Electronic Engineering, Brac University, Dhaka 1212, Bangladesh. [2]Institute of Energy, University of Dhaka, Dhaka 1000, Bangladesh.

the climate of the planet and raises average temperatures all around the globe [2]. The Asian winter monsoon has diminished due to a decrease in snow cover at mid-to-high latitudes, which has raised temperatures along the East Asian coast. Asia would likely experience significantly rising trends in mean surface temperature (0.25 to 0.34 °C per decade under RCP4.5 and 0.42 to 0.6 °C per decade under RCP8.5) [2]. The World Meteorological Organization (WMO) and the United Nations Environment Program (UNEP) developed the Intergovernmental Panel on Climate Change (IPCC) to study the causes of climate change and global warming because it is widely believed that human activity is the main contributor (UNFCCC, 2005) [3]. The extraction of greenhouse gases from air conditioners, refrigerators, and other appliances, as well as the increase in atmospheric CO_2, combustion of fossil fuels, deforestation, and other factors, all contribute to global warming. Bangladesh was ranked second among Asian nations and sixth overall in the Global Risk Index 2011 for countries that are most susceptible to natural disasters due to climate change. Bangladesh contributes only 0.3% of the emissions that cause global warming due to its low energy use. But Bangladesh is one of the worst-affected countries by the effects of global warming due to its geographic characteristics. The distinctive features that set Bangladesh's climate apart from that of other tropical regions are its high temperatures, abundant rainfall, and seasonal change [4]. Bangladesh has seen an average summer temperature of 27.5 °C over the past 30 years, which is somewhat higher than the summer average [2]. Due to the extreme poverty in this country, the problems brought on by climate change are exacerbated (ICDDR B, 2019). Action Aid's study report identified Bangladesh as the sixth most susceptible nation to famine, hurricanes, and floods [3]. Forecasting air temperatures aids meteorologists in determining the possibility in any region of the country. Air temperature is also considered an important element in evapotranspiration, which is essential for managing water supplies and agricultural operations. Many decision-making industries, including energy, transportation, and tourism, rely on accurate air temperature forecasting. Therefore, the most important component of environmental research involving functional eco-environmental systems is precisely estimating air temperature [1]. Therefore, numerous scientists and researchers over the world are attempting several investigations and creating sophisticated mathematical models to anticipate the air temperature.

Numerical-based and machine learning (ML)-based techniques are the two primary categories of weather forecasting methods used today. Numerical-based weather prediction (NWP) models include erroneous assumptions, unclear physical parameterization, and physical correlations of parameters and mechanisms of atmospheric dynamics. The model output may need to be post-processed to improve the models' effectiveness in practical applications. It raises the cost of calculation due to complicated mathematical formulas. However, ML-based techniques have gained popularity recently due to their lower processing costs and insensitivity to the multicollinearity of the input variables [5]. Hanoon et al. proposed different machine learning algorithms including Gradient Boosting Tree (GBT), Random forest (RF), Linear regression (LR), multi-layered perceptron neural network (MLP-NN), and radial basis function neural network (RBF-NN) for the prediction of air temperature [6]. The findings indicate that the MLP-NN exhibits commendable performance in forecasting daily temperature. Azamathulla presents artificial neural networks (ANNs) and gene expression programming (GEP) to predict the monthly atmospheric temperature in Tabuk, Saudi Arabia [7]. In previous research, individual ML models have typically been used to verify the predictions and demonstrate their superiority. A single forecast model is challenging to adapt to various weather parameters, even if it can increase forecast accuracy by modifying parameters and selecting features during the forecasting process. Numerous studies have demonstrated that developing ensemble and hybrid models by integrating multiple single forecast models can efficiently harness the benefits of various models and increase the precision and reliability of weather forecasts [8]. The daily temperatures in five cities throughout Belgium were predicted using a 2-layer spatiotemporal stacked LSTM model presented by Karevan et al. [9]. The results show that the spatiotemporal stacked LSTM outperformed stacked LSTM. Roy employs three deep neural networks: Multi-Layer Perceptron (MLP), Long Short Term Memory Network (LSTM), and a hybrid of Convolutional Neural Network (CNN) with LSTM [10]. Out of these models, the CNN + LSTM combination showcases the top performance, with LSTM closely trailing behind. Lee et al. utilized MLP, RNN, and CNN models to predict daily average, minimum, and maximum temperatures. They incorporated input features at a frequency greater than what had been employed in prior studies [11]. Notably, CNN, primarily utilized for processing satellite images rather than numeric weather data in temperature forecasting, surpasses the other models in performance. Mohammadi et al. developed some novel hybrid models combining autoregressive (AR), multi-layered perceptron (MLP), and autoregressive conditional heteroscedasticity (ARCH) to estimate minimum, maximum, and mean air temperatures in Northwestern Iran for both daily and monthly time scales. The research concludes that the hybrid MLP-AR models demonstrated the highest performance out of all the models tested [12]. Zhou employed a hybrid model [i.e., an artificial neural network hybridized with the powerful hetaeristic Honey Badger Algorithm (HBA-ANN)] for forecasting monthly temperatures in the hottest and coldest regions of the world

[13]. Nketiah et al. employed RNNs to construct temperature forecast models for five Chinese cities, employing five distinct model configurations. They also implemented the Ridge Regularizer (L2) during the neural network training process to prevent both overfitting and underfitting. In addition, hyperparameters were fine-tuned using the Bayesian optimization method [14].

While some studies have successfully applied ensemble and hybrid models for temperature forecasting, there may still be untapped potential in exploring different combinations of models or ensemble techniques. Further research could focus on identifying more effective ways to integrate and leverage the strengths of different models. The studies mentioned focus on specific regions, such as Belgium and Iran, and specific global extremes. There is a research gap in understanding how these models perform in a wider range of geographical contexts, including regions with different climate characteristics or extreme weather conditions. While individual studies have proposed specific lag-based schemes, there is a gap in comprehensive comparisons across various statistical schemes based on input lags. Such a comparative analysis can provide insights into the relative performance of different approaches. Additionally, previous works have not explored different statistical tests or trend analyses to identify the most appropriate approach for a given context, which may potentially lead to unreliable results.

The majority of Bangladesh is unaffected by initiatives connected to climate change, and there has been relatively limited studies based on daily temperature forecasts undertaken in this nation. The ability to effectively train policymakers and employees for mitigation and adaptation actions depends on their understanding of the nature and scope of potential climatic changes in south-eastern Bangladesh. The study uses Cox's Bazar, Bangladesh, a tropical climate case study, to forecast air temperature. The research area is distinct and located in the popular tourism eastern coastal region of the Bay of Bengal. The study applies an expansion of the well-known stacking model to complete the forecasting task. The model has been used to analyze a 20-year weather dataset that Bangladesh Meteorological Department (BMD) collected. In the current study, we proposed a perceptible stacking methodology to execute a hybrid scheme that combines the models—LR, Ridge, SVR, RF, and LGBM. They both have complimentary benefits and drawbacks which can be utilized in the stacking ensemble approach. The method chose a meta-learner and base-learners from 12 candidate models to create the stacking model's structure. By contrasting the stacking model with the individual models, the improvement in the performance of temperature forecast is demonstrated. We also compare three types of statistical schemes regarding historical time-series values of lagged days in the feature engineering stage. Also, statistical tests and trend analysis performed in this work ultimately enhanced the quality and reliability of our research findings.

2 Materials and methods

2.1 Methodology

A well-planned approach is essential for doing the investigation systematically. The elaborate framework used to conduct this study is shown in detail in Fig. 1. The methodology is mainly divided into two phases: Phase I: Preparing the data, and Phase II: Training the model.

2.1.1 Phase I: preparing the data

The phase contains gathering observed data (i.e., raw data collected form BMD), data formatting, data preprocessing, and train-test splitting. Initially, the observed weather data of 20 years (from January 1, 2001, to December 31, 2021) were obtained from BMD. Once the extraneous data had been removed, the data had been rearranged, and descriptive statistics had been calculated. In the data preprocessing stage, there are several steps i.e., missing value imputation, outlier handling, feature engineering and data normalization. After preprocessing, the dataset was divided into two sets: (i) Training set (80%) and (ii) Testing set (20%).

2.1.2 Phase II: training the model

The phase includes the stacking ensemble model set-up and model training along with out-of-fold cross-validation. The model is constructed with level-0 base-learners and level-1 meta-learner. Level-0 base learners were chosen form 11 candidate ML models based on a performance index.

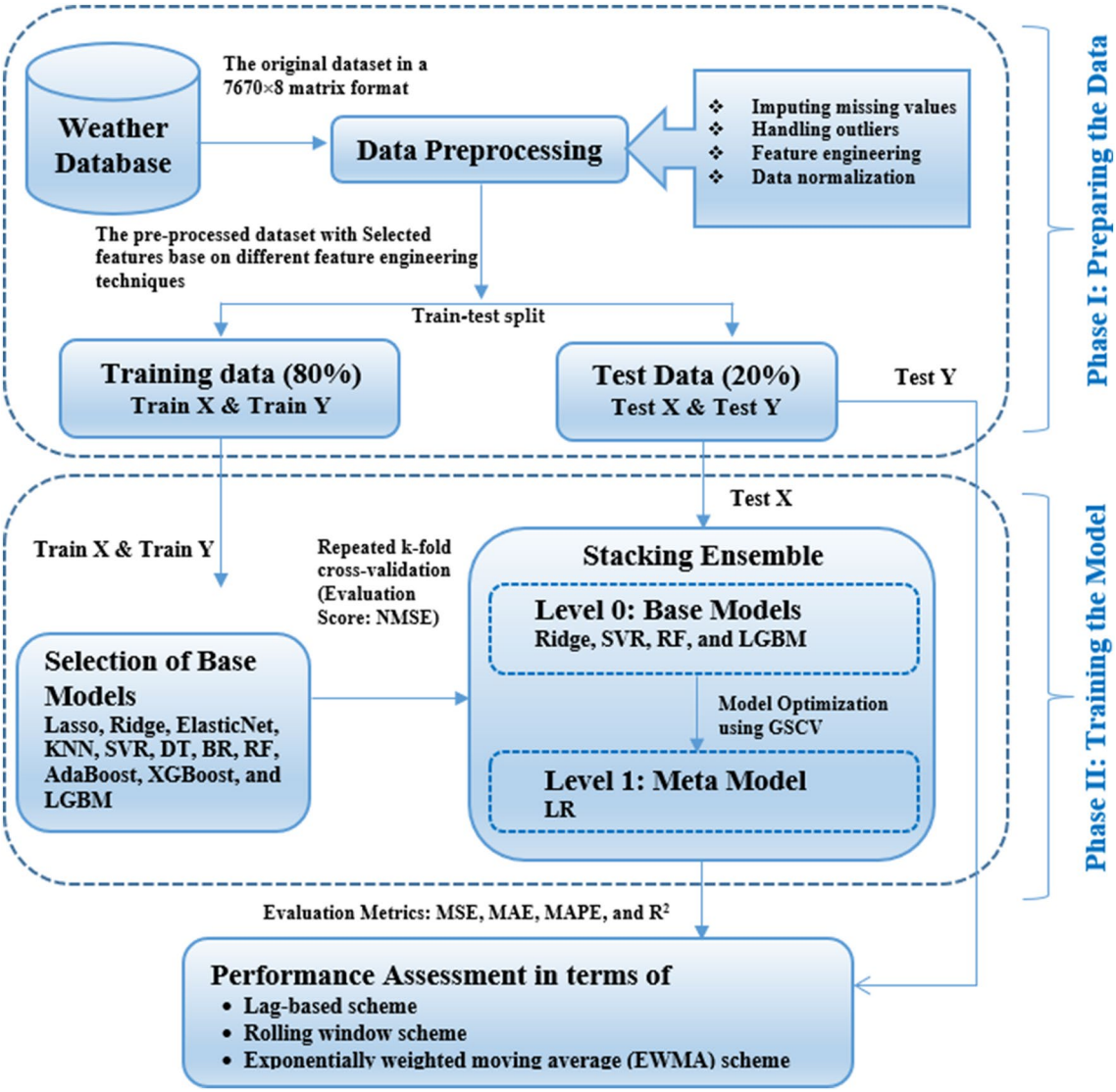

Fig. 1 Flowchart of the proposed stacking ensemble learning-based model for temperature forecasting

Lastly, the performance of the predicted values of each model was evaluated. The forecasting results were compared with the test dataset of the target variable in terms of mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). We also performed statistical tests to distinguish the significance of the performance.

2.2 Background of machine learning models

2.2.1 Stacking ensemble learning model

An ensemble approach for ML, stacked generalization was first presented by Wolpert [15]. The stacking model allows a variety of efficient models to carry out classification or regression tasks and get predictions that outperform all individual models in the ensemble. Two or more level-0 models constitute the framework of a stacking model, jointly with a level-1 model, which incorporates the predictions of the base models. Level-0 Model (Base-learner) predictions are fitted to the training set of data. The level-1 Model (Meta-learner) gains knowledge about the best methods for incorporating the

forecasts of the base models. The base-model predictions derived using data from out-of-sample are used to train the meta-learner [16].

2.2.2 Candidate machine learning models

Choosing base-learning and meta-learning combinations is a primary concern while designing stacking ensemble architecture. Stacking is suitable when several ML models have different learning skills and make distinct assumptions on the predictive modeling performance. The 12 candidate models are LR, lasso, ridge, ElasticNet, decision tree (DT), bagging regression (BR), RF, adaptive boosting (AdaBoost), LGBM, extreme gradient boosting (XGBoost), SVR, and k-nearest neighbor (KNN). Among them, all models are non-linear except LR. The meta-model is frequently straightforward, allowing for an easy interpretation of the basic model predictions. As a result, the meta-model is often a basic linear model. The base-models are selected from the remaining 11 models. Table 1 provides the characteristics of the five models selected for generating a stacking model.

2.2.3 Repeated K-fold cross-validation

Common cross-validation technique uses K-fold cross-validation (K-fold CV) for evaluating learner performance. K is any integer number. The training set's samples are divided randomly into K folds (subsets), and this procedure is repeated K times. The Kth fold of the dataset serves as the test set for all iterations, and the remaining K-1 folds serve as the training set. The process repeats until each of the K folds had served as the test set [17].

The noisy performance estimate provided by K-fold CV can make choosing a final model crucial to address the problem. An alternate approach is to execute the K-fold CV technique repeatedly and then display the mean performance of overall folds and repeats. Repeated k-fold CV is the name of this technique.

2.2.4 Grid-search cross-validation

The grid search cross-validation (GSCV) is an exhaustive search method that combines parametric search with model assessment indexes and CV techniques. The model is trained using a variety of possible parameter combinations before the grid search ultimately selects the one that yields the best results in terms of error or accuracy.

2.3 Performance metrics

During the testing phase, many metrics can be utilized to compare the performance of the models under consideration. The four performance indicators utilized for evaluation in this study are mean square error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2), shown in Table 2. In the following formula, *n* represents the number of observations, $\hat{y}_i$ represents the predicted value, and y_i represents the actual value.

Computation time: The running time of the used model is contrasted with the running time of comparison models in many articles to aid in the trade-off between the model's accuracy and computation complexity. Execution time, for instance, becomes crucial for training the model when training on continuously changing training data, such as weather conditions [20].

Statistical testing: The statistical tests were carried out in several articles to demonstrate the significance of the findings. Longer horizons and occasionally more complex models are two factors that contribute to the running time growing as the forecast horizon is increased [20]. In this research, a commonly employed nonparametric test, the Friedman test, was performed, which assesses the average rankings of benchmarking models for comparison. The Friedman test was followed by a Nemenyi post hoc analysis, which aids in identifying the most effective model for this particular dataset, ensuring accurate predictions in real-time applications of these models [5].

Trend Analysis: In weather forecasting, trend analysis plays a pivotal role in understanding long-term climate patterns and making informed predictions. In this study, the Mann–Kendall (MK) test was conducted, which is a powerful nonparametric statistical method used to detect trends or monotonic changes in time series data [21].

Table 1 A brief overview of the key aspects of five different regression models

Model	Features	Mathematical equations
LR	• LR is often used to investigate linear relationships between numerous independent variables (predictors) and the target output (dependent variable) [18]	• $\hat{y}_i = \beta_0 + \beta x_i + \epsilon_i$ x_i: input pattern y_i: output pattern $\hat{y}_i$: estimation of y_i β: regression coefficient β_0: y intercept β_n: slope of regression line ϵ: residual error
Ridge	• Ridge regression, commonly known as L2 regularization, is a kind of MLR • By incorporating a penalty component alpha, ridge regression constrains the coefficients like lasso regression [18] • In contrast to the ridge, which never reduces the value of the coefficient to zero, lasso regression frequently brings the coefficients to zero	• $\hat{\beta}^{ridge} = \mathbf{argmin} \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$ n: number of samples y_i: output pattern $\hat{y}_i$: estimation of y_i $\hat{\beta}^{ridge}$: ridge regression estimator
SVR	• SVR is a type of SVM that focuses on regression issues, intending to fit the error within a set value • The optimum fit is achieved when the number of samples between the decision boundaries is maximized • A number of kernel functions, also known as transfer functions, can be handled by the SVR. The three basic types of kernel mapping are linear, polynomial, and RBF [18]	• $\hat{y}_i = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b$ $\hat{y}_i$: estimation of y_i N: number of samples α_i, α_i^*: Lagrange multipliers $K(x_i, x_j)$: kernel function b: bias term
RF	• RF is a supervised ML algorithm that originates from Breiman's original bagging algorithm and ensemble learning • It makes use of the bootstrap resampling method to draw out several random subsets of the training data, models the decision tree for each bootstrap subset, and then averages the predictions over all decision trees [19]	• $\hat{f}_{rf}^{N} = \frac{1}{N} \sum_{n=1}^{N} T(x)$ $\hat{f}_{rf}^{N}$: predictor N: number of regression trees $T(x)$: regression Tree
LGBM	• Microsoft devised the LGBM, a gradient boosting decision tree (GBDT) method, in 2017 [5] • It makes use of two techniques: the gradient-based one-side sampling (GOSS) methodology, as well as the exclusive feature bundling (EFB) technique [8]	• $\hat{y}_i^{LG} = \sum_{p=1}^{p} f_p(x_i)$ $\hat{y}_i^{LG}$: estimation of y_i p: total number of tress f_p: trees (or models) in the ensemble

Table 2 Evaluation metric rules

Metric	Full Form	Equation
MSE	Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$
MAE	Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left\|y_i - \hat{y}_i\right\|$
MAPE	Mean absolute percentage error	$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left\|y_i - \hat{y}_i\right\|}{y_i}$
R^2	Coefficient of determination	$R^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$

3 Experimental analysis

Common ML libraries, like scikit-learn and Keras, and additional libraries, such seaborn and matplotlib, were utilized in the simulation testbed with an 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80 GHz, 1.69 GHz with 8 GB RAM, 64-bit software, and an ×64-based CPU.

3.1 Study area and data acquisition

For this study, daily weather data over the period between 1 January 2001 and 31 December 2021 of the Cox's Bazar weather station situated in the South-East part of Bangladesh was obtained from the BMD. The municipality covers 6.85 km^2 and borders the Chittagong District, the Bay of Bengal, the Bandarban District, and the Bay of Bengal [19]. Cox's Bazar has a tropical monsoon climate with an elevation of 0 feet above sea level. The average annual temperature of the district was −0.71% lower than that of Bangladesh (27.03 °C). Cox's Bazar experienced 158.42 wet days (43.4% of the time) and received about 140.97 mm (5.55 inches) of precipitation yearly. Figure 2 displays the position of the Cox's Bazar weather station, while Table 3 provides details regarding the geometric characteristics of the site. The data was provided in Excel files on a daily time scale, which contained date, temperature, maximum temperature, minimum temperature, humidity, wind speed, pressure, and rainfall.

3.2 Data handling and pre-processing

3.2.1 Formatting

The unnecessary columns, rows, and terms were deleted from the provided files (For example, Station name, and Station ID) and converted into CSV format for further use. The dataset contains 7670 rows and 7 columns, as shown in Table 4, which represents the first five records of the data obtained from January 1, 2001, to December 31, 2021. Here date, temperature, maximum temperature, lowest temperature, humidity, wind speed, pressure, and rainfall are the primary characteristics. The average, high, and low temperatures for the next day are the dependent variables of interest.

3.2.2 Checking

Table 5 shows the descriptive statistics having mean, minimum, maximum, and standard deviation. Each variable, except the wind speed, contained 7670 records (The wind speed variable had one null value). The mean for daily average temperature was 26.16 °C, the daily maximum temperature was 31.13 °C, and the daily minimum temperature was 22.16 °C. The minimum and maximum values of the average temperature were 15.50 °C and 32 °C. The measured maximum temperature values ranged from 19.70 to 39.50 °C, while the measured values of the minimum temperature ranged from 10.30 to 30 °C. According to the standard deviation, it can be seen that the distribution of total rainfall was 27.11 mm, which was the most dispersed from the mean, and the lowest one was of wind speed of 2.28 m/s.

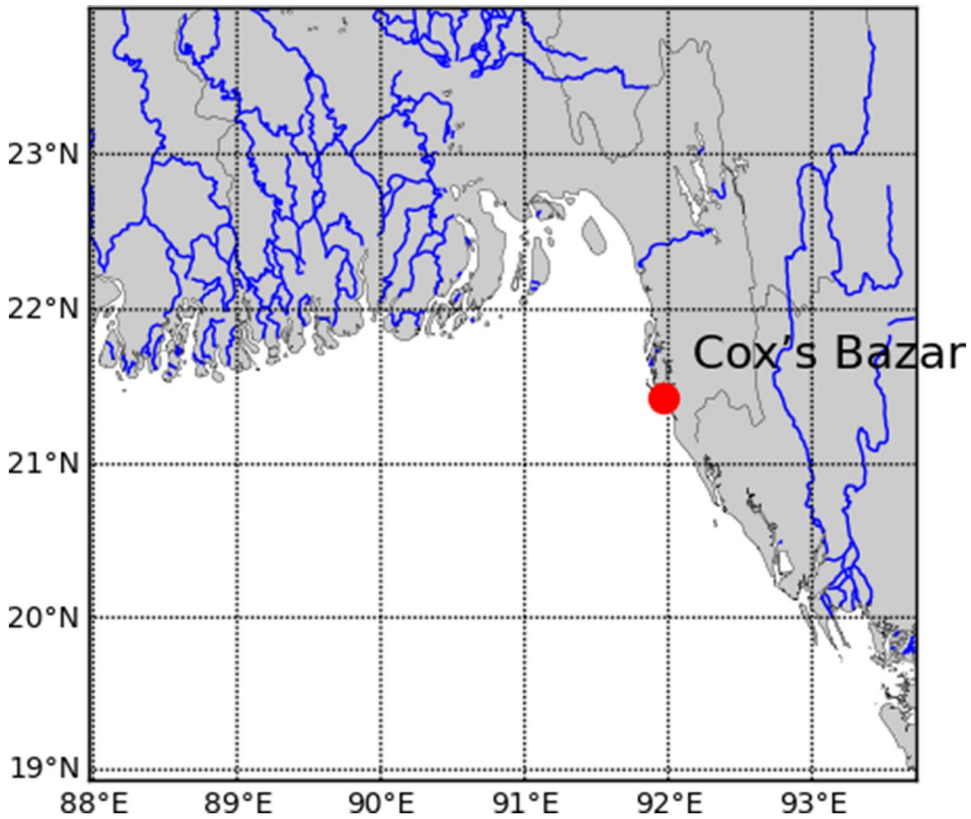

Fig. 2 Research the surrounding region and pinpoint the precise position of the Cox's Bazar weather station

Table 3 Geometric details of the selected site

Station Name	Station ID	Longitude (°E)	Latitude (°N)	Altitude (m)	Elevation from sea level (m)
Cox's Bazar	41992	91Deg. 58Mts	21Deg. 26Mts	10	2.10

Table 4 The first five records in the dataset before data preprocessing

Date	Dry-bulb temperature (°C)	Maximum temperature (°C)	Minimum temperature (°C)	Relative humidity (%)	Wind speed (m/s)	Sea level pressure (millibar)	Total rainfall (mm)
2001-01-01	21.6	29.2	16.0	75	0	1011.2	0
2001-01-02	22.5	30.2	16.5	78	1	1011.9	0
2001-01-03	22.9	30.5	18.0	76	3	1012.1	0
2001-01-04	20.9	28.0	19.0	72	8	1013.8	0
2001-01-05	19.2	25.6	15.5	65	7	1014.8	0

3.2.3 Pre-processing

Each dataset needs a pre-processing technique that ML algorithms demand. Prior to modeling, it is necessary to handle null values and outliers. Moreover, feature engineering is required to build and train better features in order to achieve effective ML.

Table 5 Descriptive Statistics of daily weather dataset (based on minimum, mean, maximum, SD, 25th, 50th, and 75th percentiles)

Features	Dry-bulb temperature (°C)	Maximum temperature (°C)	Minimum temperature (°C)	Relative humidity (%)	Wind speed (m/s)	Sea level pressure (millibar)	Total rainfall (mm)
Count	7670	7670	7670	7670	7669	7670	7670
Mean	26.16	31.13	22.61	79.58	4.15	1007.98	10.19
Std. deviation	3.12	2.56	3.92	9.01	2.28	4.70	27.11
Minimum	15.50	19.70	10.30	42	0	991	0
25%	24.10	29.50	19.50	74	3	1004.40	0
50%	27	31.60	24.30	80	4	1008.50	0
75%	28.50	33	25.50	86	5.20	1011.80	5
Maximum	32	39.50	30	100	21	1019	467

3.2.3.1 Missing value handling The data set was considered complete when the mean wind speed was used to replace the missing value.

3.2.3.2 Outliers handling The winsorization method, also known as the inter quartile range (IQR) method, was used to replace the outliers with the 25th and 75th percentiles of data for these variables [22]. According to this procedure, the following equation is given below.

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1 \tag{1}$$

where Q_3 and Q_1 represent the first quartile and third quartile. A limit for the minimum and maximum outlier values are set to cap the outliers. According to the IQR method, the normal data is defined within a range (lower limit as $\mathrm{Q}_1 - 1.5 \times \mathrm{IQR}$ and upper limit as $\mathrm{Q}_3 + 1.5 \times \mathrm{IQR}$).

3.2.3.3 Feature engineering The act of choosing, modifying, and converting unprocessed data into features that can be applied in supervised learning is known as feature engineering. We separately analyzed two complementary effects related to the feature engineering techniques: (i) The impact of the correlation-based feature selection strategy, and (ii) The inclusion of historical series values acquired in several past days.

The impact of the correlation-based feature selection strategy: Date-related features were created from the date column to extract useful information. It results the modified data with the following additional columns: month, 'day_of_month', day_of_year, week_of_year, day_of_week, year, is_month_start, is_month_end. The data already contained other attributes provided by the weather station. The process of feature selection involves lowering the number of input variables. Adding more features will make the training process difficult; however it may not always result in more accuracy [23]. The correlation heatmap was generated using the Python matplotlib and seaborn packages, which apply (1) to determine the correlation coefficient (r) between the input features [17].

$$r = \frac{\sum(x_i - x^-)\left(y_i - y^-\right)}{\sqrt{\sum(x_i - x^-)^2 \sum(y_i - y^-)^2}} \tag{2}$$

where r represents the correlation coefficient, $\mathrm{x_i}$ represents the values of x-variable in a sample, x^- represents the mean values of x-variable, $\mathrm{y_i}$ represents the values of y-variable in a sample, and y^- represents the mean values of y-variable. The correlation between every two variables is presented through the Pearson's Correlation Coefficient (PCC) heatmap in Fig. 3. The connection can be anywhere from − 1 to 1. + 1 means that there is a positive correlation, − 1 means that there is a negative correlation, and 0 means that there is no correlation. In Fig. 4 green denotes a positive, whereas red denotes a negative. The correlation magnitude increases as the color intensity does. It has been seen that the temperature has a high positive relationship with the minimum temperature. However, each of the three temperatures has a high negative correlation with the pressure. SelectKBest, a filter-based feature selection method provided by the sci-kit-learn library, has been used. It works based on the PCC between the two input variables, which can help filter out the most relevant

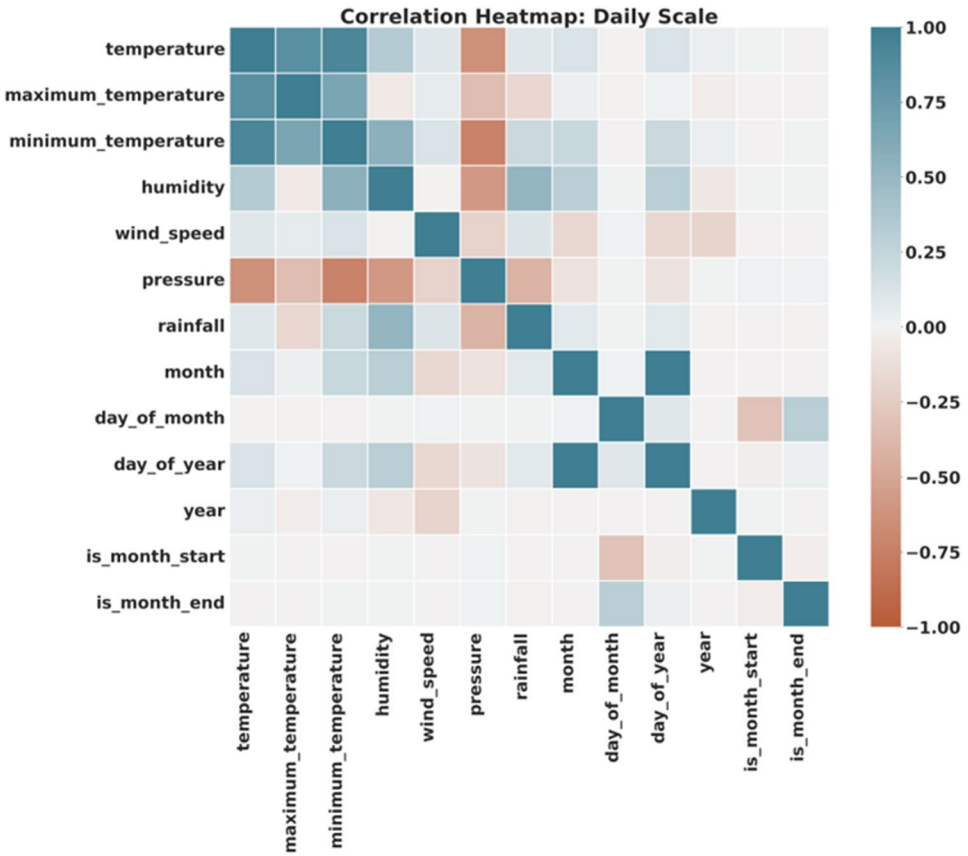

Fig. 3 Correlation among the different features visualized as a heatmap

features. A subset of the 10 most correlated features were selected and three features i.e., day_of_month, is_month_start, and is_month_end were discarded as they have the least correlation with the target variables.

The inclusion of historical series values acquired in several past days: the influence of the past value on the current value for any given observation (i.e., a time lag) was estimated using the time-lagged data from the temperature time series, as it might be advantageous to consider the significant correlated lags [23]. In Fig. 4, autocorrelation function (ACF) and partial autocorrelation function (PACF) curves with 60 lags were plotted in order to choose the appropriate input lags. Seasonality makes it is difficult to choose the best lags for the daily scale when using ACF [1]. PACF shows a high correlation with the first lag and a low correlation with the second and third lag. Therefore, compared to the ACF, the PACF is more appropriate for choosing the input lags for forecasting the small scale of time series [1]. The PACF curve shows that, the lags, out of the upper and lower bounds, can be ignored as they have lower correlations, which may result in inaccurate predictions. The time lags beginning from 1 day earlier (t−1) to 11 days earlier (t−11) are more prominent and above a 95% confidence level for each target variable. Some feature engineering methods extend beyond simply adding raw lagged values by computing some statistical values based on past values. Among them, rolling window and exponentially weighted moving average (EWMA) schemes are generally effective for improving the performance of time series forecasting models; but using appropriate method is important to avoid overfitting.

Rolling mean features involve calculating the mean of the time series data over a specified window size and using the mean values as input features for the forecasting model. The number of lagged values determines the window size. Exponentially weighted mean features are similar to rolling mean features; but instead of using a fixed window size to calculate the mean, they use an exponentially decaying weighting scheme. The decay factor in the weighting scheme can be determined to control the importance given to recent observations. Table 6 presents the selected features applied in this study.

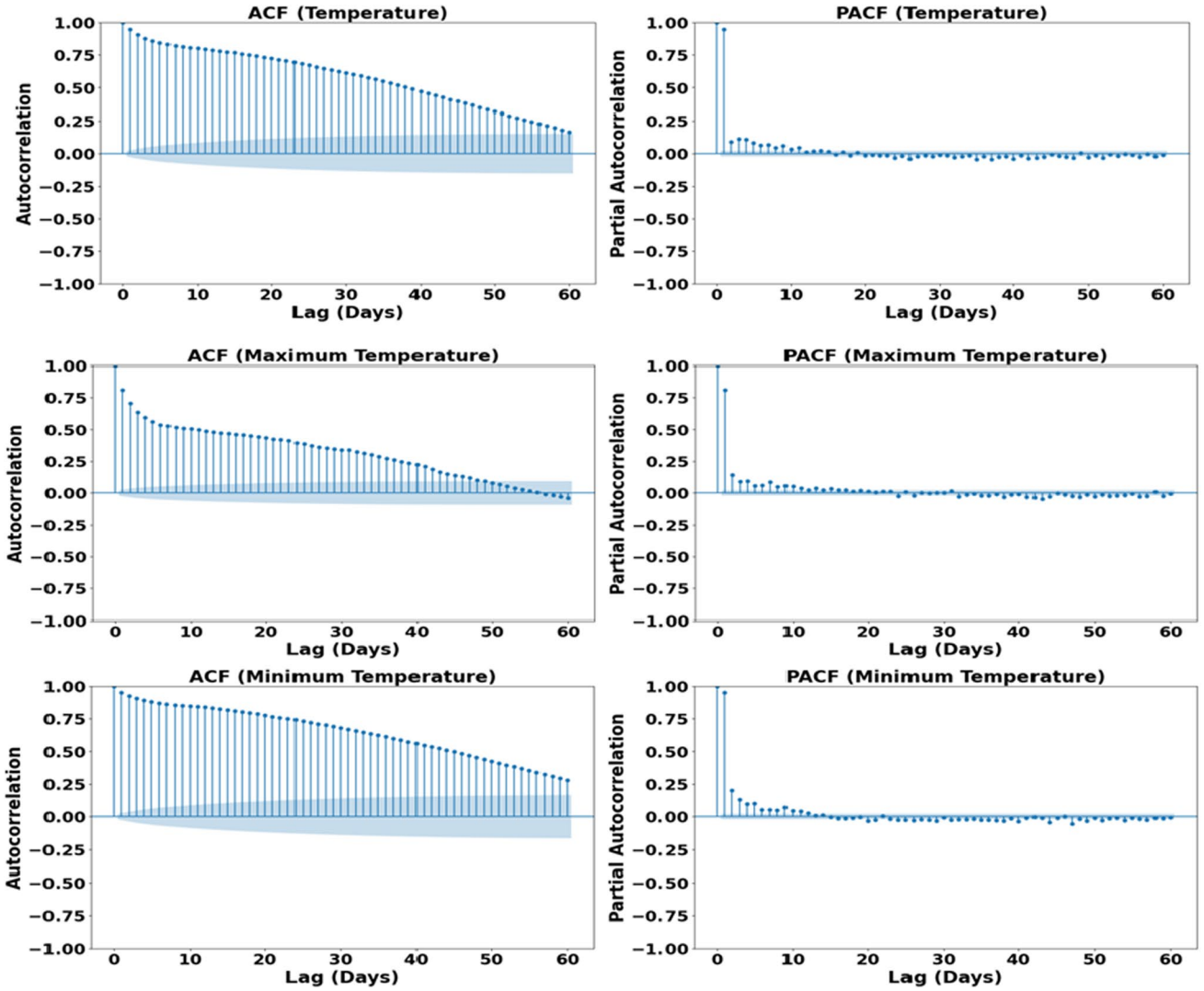

Fig. 4 Temperature, maximum temperature, and lowest temperature, as well as their associated ACF and PACF curves with regard to 60 delays

Table 6 Selected features based on feature engineering

Feature engineering techniques	Selected features
Domain-specific features	Temperature, maximum temperature, minimum temperature, humidity, rainfall, pressure, and wind speed
Date-related features	month, day_of_year, year
Time-lag features	Input lags (Lag 1, lag 2, lag 3, ……, lag 11)/Rolling window features (Rolling window mean of input lags)/ exponential weighting features (exponentially weighted mean of input lags)

3.2.4 Normalization

Data normalization is frequently employed in ML techniques to lessen the impact of the variety of data. In this research, for data normalization, robust scaling is used [24]. It uses median and interquartile range (IQR) to scale input values. Robust scaling is resistant to the negative effects of outliers. The formula is as follows:

$$\text{Scaled Value} = \frac{\text{Original Value} - \text{Input Median}}{\text{Input IQR}} \tag{3}$$

4 Model set-up

4.1 Data partition

A training set and a testing set were created by dividing the dataset in advance of the training process. Eighty percent (80%) of the dataset was used for training, while the other twenty percent (20%) was used for testing. The model can be constructed and fitted to the available data with the aid of the training set. Estimating the model's efficacy on new data (data not used to train the model) was made easier with the help of the testing set.

4.2 Stacking model construction

Each of the possible models adheres to a slightly different set of rules for fitting data sets, and each has its own set of benefits when working with certain classes of variables.

4.2.1 Model selection

In this study the linear regression (LR) was chosen as the meta-learner. Since linear regression does not need a second hyper-parameter to be adjusted and has weights that can be interpreted to indicate base-learner significance, it is frequently used in regression for the meta-learner. We employed 11 potential models with the default hyper-parameters to compare their performance on the training dataset using repeated tenfold CV with several iterations of 3. The negative mean squared error (NMSE) is considered as the selection score, given in Table 7. The base-learners for the stacking-model were picked from the four models that did the best. Table 7 shows that LGBM had the lowest average CV error, with an MSE of 0.912, 0.895, and 0.897, respectively. The other three best-performed models were Ridge, SVR, and RF. Therefore, LGBM, Ridge, SVR, and RF were selected as base-learners.

The selected models can characterize the base-learners that should possess diversity. The radial basis function (RBF) kernel was proved efficient in prior meteorological investigations of SVR application [16]. For the maximum learning problems, RF has approximately the same error rate as the other methods and is less prone to overfitting [8]. LGBM offers effective parallel training with the advantages of a quick training speed, low memory consumption, and the capacity to process massive volumes of data, partially addressing the drawbacks of conventional models [15]. Ridge performs well when a data set has multicollinearity (correlations between predictor variables).

Table 7 MSE (°C) of the 11 candidate models obtained by repeated K-fold CV (selected models are bolded)

Algorithm	Lag-based scheme			Rolling window scheme			EWMA scheme		
	T	Max T	Min T	T	Max T	Min T	T	Max T	Min T
Lasso	3.030	4.183	3.370	2.959	4.159	3.378	3.010	4.103	3.358
Ridge	**0.918**	**1.947**	**0.724**	**0.915**	**1.942**	**0.725**	**0.916**	**1.941**	**0.724**
ELNET	2.254	3.268	2.592	2.625	3.258	3.147	2.646	3.236	3.166
KNN	1.231	2.488	1.297	1.189	2.333	1.185	1.192	2.316	1.197
SVR	**0.935**	**1.995**	**0.778**	**0.924**	**1.963**	**0.765**	**0.930**	**1.967**	**0.764**
DT	1.825	3.991	1.586	1.798	3.848	1.510	1.823	3.844	1.517
BR	1.012	2.126	0.861	1.004	2.164	0.843	1.006	2.129	0.843
RF	**0.922**	**1.944**	**0.779**	**0.925**	**1.954**	**0.773**	**0.926**	**1.951**	**0.766**
ADB	1.327	3.485	1.237	1.289	3.472	1.265	1.324	3.462	1.220
XGB	1.012	2.170	0.860	0.983	2.108	0.823	0.977	2.132	0.821
LGBM	**0.912**	**1.941**	**0.759**	**0.895**	**1.917**	**0.739**	**0.897**	**1.927**	**0.739**

T Temperature, *Max T* maximum temperature, *Min T* minimum temperature

Table 8 Hyper-parameters of the base-learners and range of grid search values used in the study

Algorithm	Hyper-parameters	Grid search values
Ridge	alpha	0, 1, 2, 3, 4, 5, …., 100
SVR	cost	0.1, 1, 10
	gamma	0.1, 1, 10
	epsilon	0.1, 1, 10
RF	max_depth	1, 3, 5, 7, 9, 11
	max_features	3, 5, 15
	n_estimators	50, 100, 150, 200, 250, …., 500
	min_samples_split	2, 5, 8
LGBM	learning_rate	0.01, 0.05, 0.1
	max_depth	1, 3, 5, 7, 9, 11
	n_estimators	50, 100, 150, 200, 250, …., 500
	subsample	0.5, 0.7, 0.9, 1

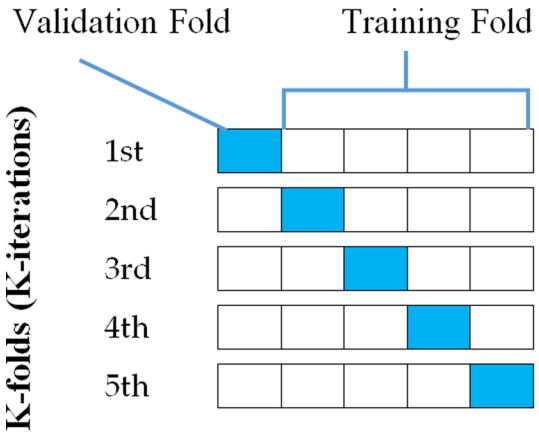

Fig. 5 Fivefold CV for training and testing set division to fit and evaluate the stacking model

4.2.2 Hyper-parameter optimization

Since the performance of the base-learners is affected by a number of hyper-parameters, the grid search CV was used to find the optimal combination of hyper-parameters within the allowed range based on the results of prior studies [25]. It used negative mean squared error (NMSE) and tenfold CV to rank the effectiveness of the many permutations of the model's hyper-parameters that were provided. Then, the optimized models were applied as base-learners of the stacking model. The gird search values used for each base model are presented in Table 8.

4.2.3 Stacking model implementation

The proposed model implemented the stacking regression with fivefold CV, which used the concept of out-of-fold predictions to prepare the input data for the level-1 meta-learner. The training dataset was divided into five folds. In five successive rounds, four folds were used to fit the level-0 base-learners. In each iteration the optimized level-0 models were applied to the remaining one subset. The resulting predictions were then stacked and provided to the level-1 meta-learner. Figure 5 shows the fivefold CV for training and testing set division to fit and evaluate the stacking model.

5 Result and discussion

5.1 Inter-comparison of model performances

The forecasting was evaluated and compared to that of the optimized base models in terms of MSE, MAE, MAPE, and R^2 for the three target variables. In addition, we performed an experimental analysis on the feature engineering statistical

Table 9 Evaluation results of forecasting daily average temperature, maximum temperature and minimum temperature using lag-based scheme (the best results are boldfaced)

Variable	Algorithm	Training Phase				Testing Phase			
		MSE (°C)	MAE (°C)	MAPE (%)	R^2	MSE (°C)	MAE (°C)	MAPE (%)	R^2
Average temperature	LR	0.909	0.713	2.783	0.905	0.800	0.679	2.635	0.919
	Ridge	0.909	0.713	2.784	0.905	0.801	0.679	2.636	0.919
	SVR	0.830	0.643	2.530	0.914	0.887	0.717	2.791	0.911
	RF	**0.546**	**0.572**	**2.212**	**0.943**	0.781	0.679	2.622	0.921
	LGBM	0.777	0.666	2.595	0.919	0.784	0.683	2.638	0.921
	Stacking	0.715	0.635	2.471	0.925	**0.755**	**0.662**	**2.560**	**0.924**
Maximum temperature	LR	1.929	0.984	3.265	0.693	1.907	1.044	3.451	0.727
	Ridge	1.929	0.984	3.265	0.693	1.910	1.044	3.453	0.726
	SVR	1.725	0.944	3.136	0.726	2.219	1.134	3.769	0.682
	RF	**1.074**	**0.769**	**2.533**	**0.829**	1.933	1.035	3.439	0.723
	LGBM	1.438	0.857	2.838	0.771	1.912	1.037	3.441	0.726
	Stacking	1.386	0.849	2.808	0.780	**1.859**	**1.024**	**3.394**	**0.733**
Minimum temperature	LR	0.719	0.615	2.884	0.953	0.563	0.548	2.569	0.963
	Ridge	0.719	0.615	2.884	0.953	0.564	0.564	2.570	0.963
	SVR	0.512	0.463	2.168	0.967	0.793	0.658	3.127	0.948
	RF	**0.425**	**0.494**	**2.281**	**0.972**	0.618	0.584	2.732	0.959
	LGBM	0.507	0.522	2.441	0.967	0.583	0.566	2.662	0.962
	Stacking	0.561	0.546	2.544	0.964	**0.535**	**0.537**	**2.512**	**0.965**

Table 10 Evaluation results of forecasting daily temperature, maximum temperature and minimum temperature using rolling window scheme (the best results are boldfaced)

Variable	Algorithm	Training Phase				Testing Phase			
		MSE (°C)	MAE (°C)	MAPE (%)	R^2	MSE (°C)	MAE (°C)	MAPE (%)	R^2
Average temperature	LR	0.911	0.713	2.784	0.905	0.796	0.677	2.626	0.920
	Ridge	0.911	0.713	2.785	0.905	0.796	0.677	2.627	0.920
	SVR	0.782	0.689	2.674	0.918	0.830	0.705	2.707	0.916
	RF	**0.568**	**0.578**	**2.231**	**0.941**	0.788	0.684	2.635	0.921
	LGBM	0.790	0.669	2.604	0.918	0.781	0.681	2.627	0.921
	Stacking	0.686	0.623	2.421	0.928	**0.750**	**0.657**	**2.538**	**0.924**
Maximum temperature	LR	1.934	0.985	3.267	0.693	1.907	1.042	3.444	0.727
	Ridge	1.934	0.985	3.267	0.693	1.910	1.042	3.446	0.726
	SVR	1.820	0.959	3.186	0.711	2.038	1.074	3.563	0.708
	RF	**1.229**	**0.805**	**2.658**	**0.805**	1.883	1.025	3.407	0.730
	LGBM	1.688	0.923	3.057	0.732	1.881	1.034	3.422	0.730
	Stacking	1.490	0.873	2.889	0.763	**1.839**	**1.022**	**3.381**	**0.736**
Minimum temperature	LR	0.722	0.616	2.888	0.953	0.562	0.546	2.561	0.963
	Ridge	0.722	0.616	2.888	0.953	0.563	0.546	2.561	0.963
	SVR	0.619	0.538	2.535	0.960	0.704	0.615	2.913	0.954
	RF	**0.437**	**0.497**	**2.295**	**0.972**	0.615	0.582	2.732	0.960
	LGBM	0.570	0.551	2.579	0.963	0.590	0.566	2.659	0.961
	Stacking	0.582	0.555	2.588	0.963	**0.535**	**0.537**	**2.515**	**0.965**

Table 11 Evaluation results of forecasting daily temperature, maximum temperature and minimum temperature using EWMA scheme (the best results are boldfaced)

Variable	Algorithm	Training Phase				Testing Phase			
		MSE (°C)	MAE (°C)	MAPE (%)	R^2	MSE (°C)	MAE (°C)	MAPE (%)	R^2
Average temperature	LR	0.912	0.714	2.791	0.905	0.792	0.674	2.614	0.920
	Ridge	0.912	0.715	2.792	0.905	0.792	0.674	2.615	0.920
	SVR	0.785	0.691	2.681	0.919	0.827	0.703	2.700	0.917
	RF	**0.597**	**0.591**	**2.289**	**0.938**	0.782	0.679	2.616	0.921
	LGBM	0.790	0.669	2.611	0.918	0.777	0.679	2.617	0.922
	Stacking	0.701	0.627	2.441	0.927	**0.749**	**0.656**	**2.537**	**0.925**
Maximum temperature	LR	1.931	0.984	3.266	0.694	1.906	1.040	3.440	0.727
	Ridge	1.931	0.984	3.266	0.694	1.909	1.041	3.442	0.726
	SVR	1.823	0.959	3.190	0.711	2.035	1.074	3.562	0.708
	RF	**1.219**	**0.804**	**2.652**	**0.807**	1.892	1.030	3.424	0.729
	LGBM	1.429	0.857	2.837	0.774	1.895	1.040	3.446	0.728
	Stacking	1.449	0.862	2.851	0.770	**1.839**	**1.025**	**3.389**	**0.736**
Minimum temperature	LR	0.721	0.615	2.889	0.953	0.561	0.545	2.555	0.963
	Ridge	0.721	0.615	2.889	0.953	0.561	0.545	2.554	0.963
	SVR	0.618	0.537	2.535	0.960	0.699	0.613	2.902	0.954
	RF	**0.436**	**0.496**	**2.294**	**0.972**	0.617	0.582	2.729	0.959
	LGBM	0.574	0.552	2.589	0.963	0.586	0.562	2.641	0.961
	Stacking	0.581	0.555	2.589	0.962	**0.535**	**0.536**	**2.510**	**0.965**

schemes—lag-based, rolling window, and EWMA. The experimental outcomes for each variable for the training and testing phases are presented in Tables 9, 10, 11.

All models demonstrate satisfactory performance in predicting average, maximum, and minimum temperatures during the training phase. Notably, the RF model outperforms others in predicting all three temperature variables (average, maximum, and minimum). On the contrary, the performance of the Ridge model was inferior compared to the other models. For average temperature (T) prediction during the training phase, the RF model exhibits the best performance, with MSE values of 0.546, 0.568, and 0.597 for lag-based, rolling window, and EWMA schemes, respectively. The RF model also yields MAE values ranging from 0.572 to 0.591, MAPE values ranging from 2.212 to 2.231%, and R^2 values ranging from 0.938 to 0.943. Following closely is the stacking model, with MSE ranging from 0.686 to 0.715, MAE ranging from 0.623 to 0.635, MAPE ranging from 2.421 to 2.471, and R^2 ranging from 0.925 to 0.928. In contrast, Ridge models exhibit the least favorable performance, with MSE ranging from 0.909 to 0.912, MAE ranging from 0.713 to 0.715, MAPE ranging from 2.784 to 2.792, and R^2 values of 0.905.

In the case of forecasting maximum temperature (Tmax), the RF model outperforms the others, closely followed by the stacking model in most instances. The RF model achieves MSE values ranging from 1.074 to 1.229, mean absolute error (MAE) values ranging from 0.769 to 0.805, MAPE values ranging from 2.533% to 2.658%, and R^2 values ranging from 0.829 to 0.805. For minimum temperature (Tmin) prediction, the RF model achieves MSE values ranging from 0.425 to 0.437, MAE values ranging from 0.494 to 0.497, MAPE values ranging from 2.281% to 2.295%, and a R^2 value of 0.972. The RF model's performance is followed by LGBM and then the stacking model across all three statistical schemes. Additionally, in the training phase, the lag-based scheme tends to enhance the accuracy of predictions for T, Tmax, and Tmin for all models, with the exception of LGBM's performance in forecasting Tmax using the EWMA scheme.

All models demonstrate strong predictive capabilities, particularly in forecasting T and Tmin during the training phase. However, it's important to emphasize that assessing model performance based on the testing dataset is of paramount importance. In the training phase, models are exposed to complete data, including input features and target values, which can potentially lead to overfitting. Therefore, excluding the evaluation of models in the testing phase could yield misleading results. It is worth noting that during testing, models only receive input features, making the forecasting accuracy more dependable compared to the training phase.

In the testing phase, the stacking model outperforms all other models by a certain margin, as measured by a variety of performance metrics applied to all target variables. SVR stands out as the poorest performer across all

three temperature forecasting schemes. In forecasting the average temperature, the stacking model has a reduced MSE range between 0.749 and 0.755 °C. The MAE values of the stacking model were 0.656 °C–0.662 °C, whereas MAPE values were 2.538%–2.560%. For the stacking model, the R-squared (R^2) value is 0.925 when using the EWMA scheme, while it stands at 0.924 for both the lag and rolling window schemes. Based on the result, the LGBM model comes in second, followed by the RF model. The performance of SVR is comparatively worse than others in respect of all evaluation metrics, with MSE ranging from 0.827 to 0.887, MAE ranging from 0.703 to 0.717, MAPE ranging from 2.700 to 2.791%, and R^2 ranging from 0.911 to 0.917.

In the context of predicting Tmax, the stacking model exhibits MSE values between 1.839 and 1.859, MAE values between 1.022 and 1.025, mean absolute percentage error (MAPE) values from 3.381 to 3.394%, and R^2 values from 0.733 to 0.736. For Tmin, all models perform adequately, with the stacking model displaying superior performance. It attains an MSE value of 0.0535 and a R^2 value of 0.965 across all schemes. Additionally, the model achieves MAE values ranging from 0.536 to 0.537 and MAPE values ranging from 2.510% to 2.515%.

However, the RF model outperforms other models in the training phase (Tables 9, 10, 11). This indicates a tendency towards overfitting. It means that the RF model captures noise or random variations specific to the training data, which may not translate well to new, unseen data. On the contrary, the stacking model exhibits a minimal disparity between its performance in the training and testing phases. This indicates that the model was adeptly shielded against overfitting. Stacking models, by design, possess the potential to generalize more effectively compared to individual models. In this framework, a meta-learner is integrated to learn how to combine the predictions of base models, ultimately enhancing overall performance on unseen data. In terms of forecasting T, the stacking model demonstrated MSE improvement rates of 13.2%, 8%, and 7.8% across the three schemes compared to SVR. Additionally, the MAE improvement rates stood at 5.5%, 4.8%, and 4.7%. When it comes to forecasting Tmax, the MSE improvement rates ranged from 19.6 to 36%, and the MAE increment rates ranged from 4.9 to 11%. Similarly, for Tmin, the MSE improvement rates varied from 16.4 to 25.8%, and the MAE increment rates ranged from 7.7 to 12.1%.

To visually represent the predictions made by the algorithm, Fig. 6 displays both the observed data from BMD and the predicted values generated by the stacking model for the most recent 365 samples. Subgroups (a), (b), and (c) correspond to the three schemes: lag-based, rolling window, and EWMA. The graphs clearly illustrate the stacking model's accurate portrayal of time-series variations when compared to the observed data. This reaffirms the model's capacity to generalize the actual temperature profile. Additionally, Figure (b) highlights that when forecasting maximum temperature, the model's performance is relatively lower. This could be attributed to potential challenges in capturing more pronounced temporal trends or patterns specific to maximum temperature data. From Figs. 6a–c, it can be seen that the three statistical schemes exhibit no significant differences in their performance when forecasting the three temperatures (T, Tmax, and Tmin). A comparison of the existing machine learning and other approaches to the prediction of temperature is summarized in Table 12.

The scatter plots of the stacking model and base models presented in Fig. 7 with sub-groups (a), (b), and (c) for three different schemes. It revealed that the maximum clustered points were close to the diagonal line in case of average temperature and minimum temperature forecasting which confirms that the prediction results of these two weather parameters were highly coincidental with the experimental data. The maximum temperature-predicted data points were comparatively scattered. R^2 values of the stacking model and the base models mentioned in Table 11 also offered the same indication. The R^2 values of the models were 0.948 or higher in minimum temperature forecasting. R^2 more than 0.75 was a very good fit, and R^2 equivalent 0.64 to 0.74 was a good fit [6]. All ML methods perform well with accuracy (more than 68.2%). Among them, the stacking model has the best performance with accuracy more than 96.5%. There was no significant difference in evaluation metrics with respect to three feature engineering schemes.

Figure 8 visually represents the prediction errors during the testing phase using a probability density curve. This curve is constructed using the Gaussian kernel density function, ensuring that the errors follow a normal distribution. For each target variable, the graphic depicted a distinct error distribution over the range of schemes and models. The figure demonstrated that the probability density curve derived using the stacking model was a superior match, which was also verified by the standard deviation values provided in Table 13. The stacking model consistently exhibits a lower standard deviation of predicted data compared to the other models in all scenarios. Specifically, for the EWMA scheme, the standard deviations are 0.866, 1.355, and 0.731 for T, Tmax, and Tmin, respectively.

It is noteworthy to mention that the choice between statistical schemes like EWMA (Exponentially Weighted Moving Average), rolling window, and lag values depends on the specific characteristics of the data. Here, the EWMA approach considerably enhances the predictive capability of the stacking model, particularly in the forecasting of T and Tmin. Optimal accuracy is achieved in forecasting the Tmax when the rolling window scheme is employed. It can be said that

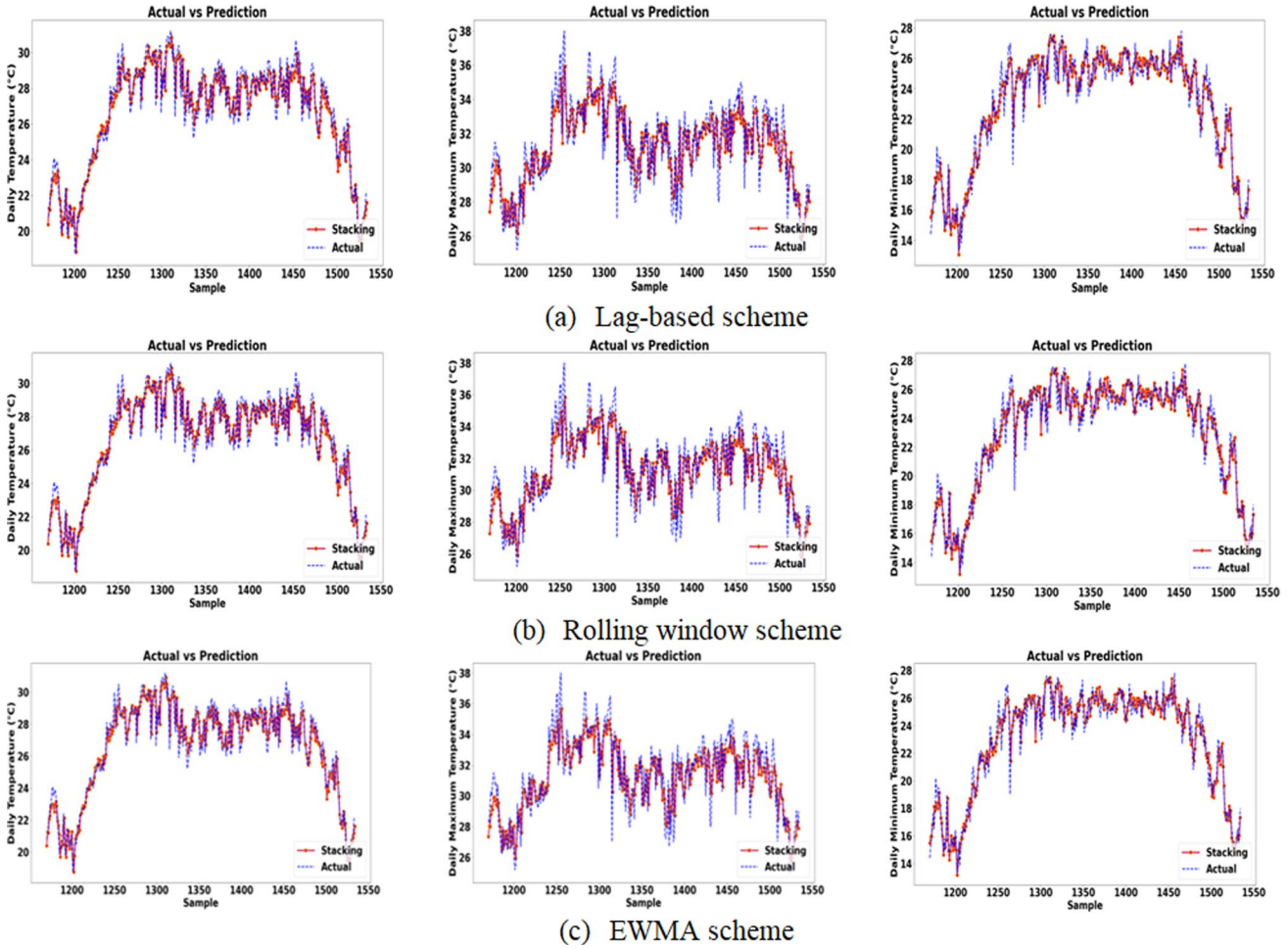

Fig. 6 Forecasting of temperature, maximum temperature, and minimum temperature using stacking model for three feature engineering statistical schemes: **a** lag-based, **b** rolling window, and **c** EWMA

the rolling window and EWMA schemes might be more adaptable to changes or variations in the data patterns. They may have a smoother response to shifts in the underlying trends, which can be beneficial when dealing with unseen data.

5.2 Computation time analysis

As depicted in Table 14, the Ridge model exhibited the shortest computational time. However, the RF model, owing to its multitude of finely tuned hyperparameters, displayed a moderately lower level of prediction accuracy. In contrast, the LGBM model demonstrated significantly shorter execution times compared to its counterparts. The stacking model, a key recommendation of this research, necessitated a relatively longer computational time. Nonetheless, this increased duration was accompanied by the model achieving the highest level of performance among all evaluated methods. This highlights the trade-off between computational efficiency and predictive accuracy, underscoring the stacking model's potential as a robust forecasting tool.

5.3 Statistical test analysis

The non-parametric Friedman test has been conducted due to the shortage of comparison methods. The Friedman test, a two-way statistical assessment of rank variance, assumes that all machine learning algorithms perform equally under the null hypothesis (H0) in relation to performance metrics like MSE, MAE, MAPE, and R^2. Conversely, the alternative hypothesis (HA), suggests that at least one approach exhibits a notable difference in performance. Friedman test results on the performance metrics are given in Table 15. Based on the outcomes presented in Table 10,

Table 12 Performance comparison of the proposed model with the existing models

References	City/country	Prediction Models	Best Model	Performance Metrics		
				Tavg	Tmax	Tmin
Alomar et al. [1]	North Dakota, North America	ARIMA, RT, SVR, GBR, QRT, RF	SVR	RMSE = 3.592 MAE = 2.745 R^2 = 0.964 U = 0.127	–	–
Apaydın et al. [5]	Seoul, South Korea	Ridge, Lasso, ENet, SGD, KNN, DT, RF, ANN, SVR, ADB, LGBM	XGB, LGBM	–	R^2 = 0.93 (2013) R^2 = 0.94 (2014) R^2 = 0.92 (2015) R^2 = 0.97 (2016) R^2 = 0.93 (2017)	R^2 = 0.91 (2013) R^2 = 0.91 (2014) R^2 = 0.93 (2015) R^2 = 0.96 (2016) R^2 = 0.94 (2017)
Karevan and Suykens[9]	Brussels, Antwerp, Liege, Amsterdam and Eindhoven	Stacked LSTM, ST stacked LSTM	ST Stacked LSTM	–	MAE = 1.15 MSE = 2.48	MAE = 1.43 MSE = 3.64
Roy [10]	John F. Kennedy International Airport, New York	MLP, LSTM, CNN, CNN + LSTM	CNN + LSTM	MAPE = 2.58	–	–
Lee et al. [11]	Seoul, Daegwalleong, and Seongsan	MLP, RNN, CNN	CNN	MAE = 1.308 (Seoul) MAE = 1.700 (Daegwallyeong) MAE = 1.109 (Seongsan)	MAE = 2.053 (Seoul) MAE = 2.331 (Daegwallyeong) MAE = 1.671 (Seongsan)	MAE = 1.246 (Seoul) MAE = 1.700 (Daegwallyeong) MAE = 1.199 (Seongsan)
Hanoon et al. [6]	Kula Terengganu, Malaysia	GBT, RF, LR, ANN, MLP-NN, RBF-NN	MLP-NN	MAE = 0.018 RMSE = 0.026 R^2 = 0.713	–	–
Huang et al. [26]	Guangxi, China	RNN combined with rough set, stepwise regression	RNN combined with rough set	–	MAE = 1.98	MAE = 1.40
Azamathulla et al. [7]	Tabuk, Saudi Arabia	ANN, GEP	GEP	RMSE = 0.44 R^2 = 0.91		
Mohammadi et al. [12]	Northwestern Iran	MLP, MLP-AR, MLP-AR-ARCH	MLP-AR(14), MLP-AR(12), AR(15)-ARCH	RMSE = 0.262 MAE = 0.194 NRMSE = 2.155	RMSE = 0.364 MAE = 0.277 NRMSE = 1.911	RMSE = 0.253 MAE = 0.140 NRMSE = 4.801
This study	Cox's Bazar, Bangladesh	LR, Ridge, SVR, RF, LGBM, Stacking	Stacking	MSE = 0.749 MAE = 0.656 MAPE = 2.537 R^2 = 0.925	MSE = 1.839 MAE = 1.022 MAPE = 3.381 R^2 = 0.736	MSE = 0.535 MAE = 0.536 MAPE = 2.510 R^2 = 0.965

ARIMA Autoregressive integrated moving average, *LR* linear regression, *RT* regression tree, *SVR* support vector regression, *GBR* gradient boosting regression, *GBT* gradient boosting tree, *QRT* quantile regression tree, *RF* random forest, *MLP* multilayer perceptron, *NN* neural network, *RBF* radial basis function, *SGD* stochastic gradient descent, *KNN* K-nearest neighbor, *DT* decision tree, *ANN* artificial neural network, *GEP* gene expression programming, *LGBM* light gradient boosting machine, *XGB* extreme gradient boosting, *ADB* adaptive boosting, *ENet* Elastic-Net, *ST Stacked LSTM* spatio-temporal stacked LSTM, *LSTM* long short term memory, *CNN* convolutional neural network, *RNN* recurrent neural network, *RMSE* root mean square error, *MAE* mean absolute error, *U* U-statistics, R^2 correlation coefficient

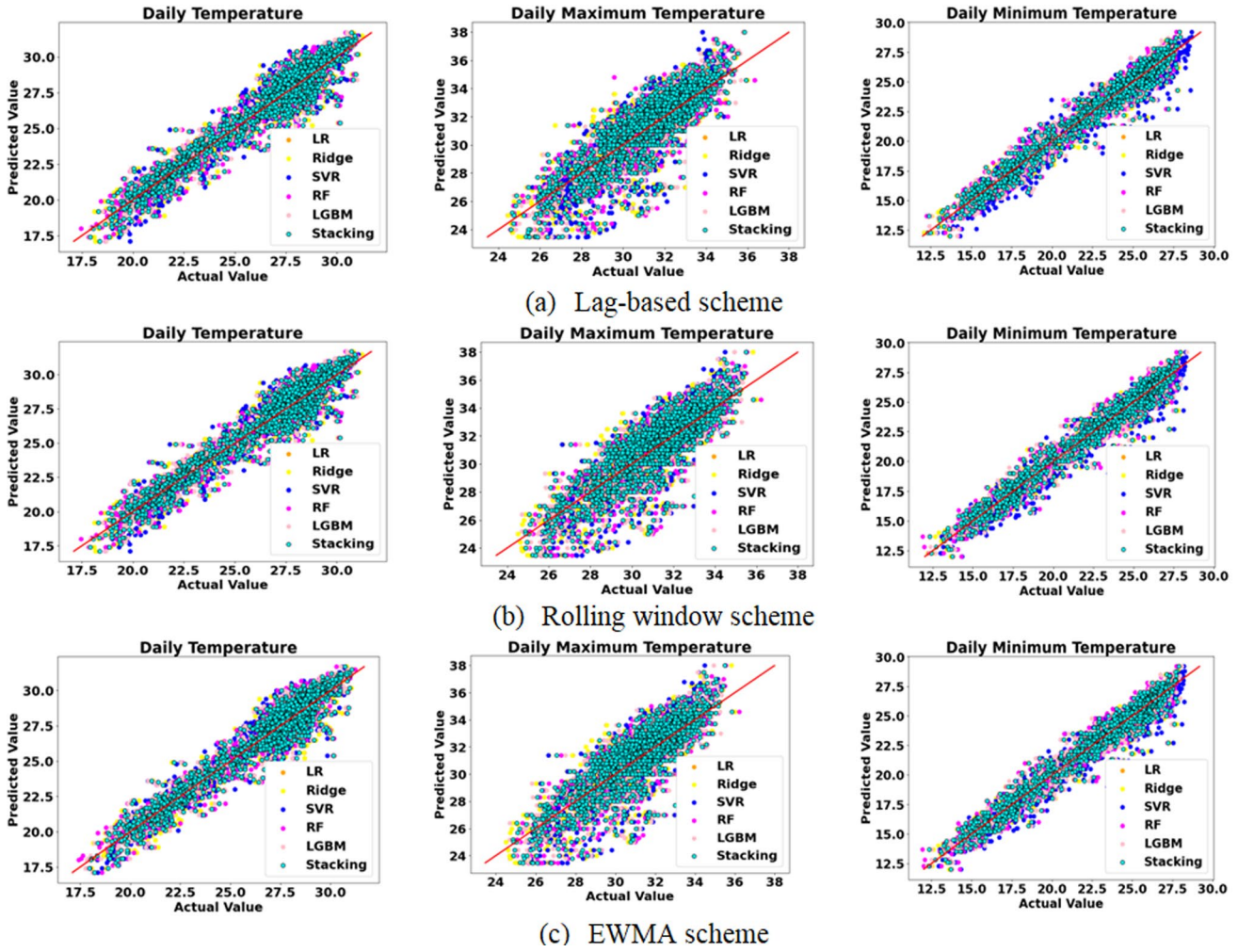

Fig. 7 Scatter plots showing the association between actual value and predicted value of the target variables obtained by the stacking model for three feature engineering statistical schemes: **a** lag-based, **b** rolling window, and **c** EWMA

the Friedman test dismisses the null hypothesis for all temperature conditions. This is indicated by the p-values being below the designated significance level ($\alpha = 0.05$) and the FF values surpassing the critical value of 19.675. Consequently, the null hypothesis of the Friedman test is rejected, providing evidence of a substantial distinction between the models under comparison. As the Friedman test is significant, the average ranks are determined and presented in Table 16.

The Friedman test, however, was not enough for making such a distinction. Therefore, a Nemenyi test was run to compare the six models' accuracy levels and determine whether or not there was statistically significant variance. The crucial difference (CD) for the Nemenyi test was 2.176. This means that there was a statistically significant difference in performance across models if the difference between their average rankings was more than 2.176. Figure 9 shows the Nemenyi test's CD diagram for all dependent variables.

The diagram illustrates a clear distinction between the six models, categorized based on their average ranking. For temperature prediction, the top-performing methods fall into two groups: stacking, LGBM, and RF form one category, while LR, Ridge, and SVR comprise the other. In terms of maximum temperature forecasting, the superior models are categorized into stacking, RF, and LGBM on the one hand, and LR, Ridge, and SVR on the other. Finally, for minimum temperature prediction, the standout models are categorized into two groups. One category contains stacking, LR, and Ridge, and the other category contains LGBM, RF, and SVR.

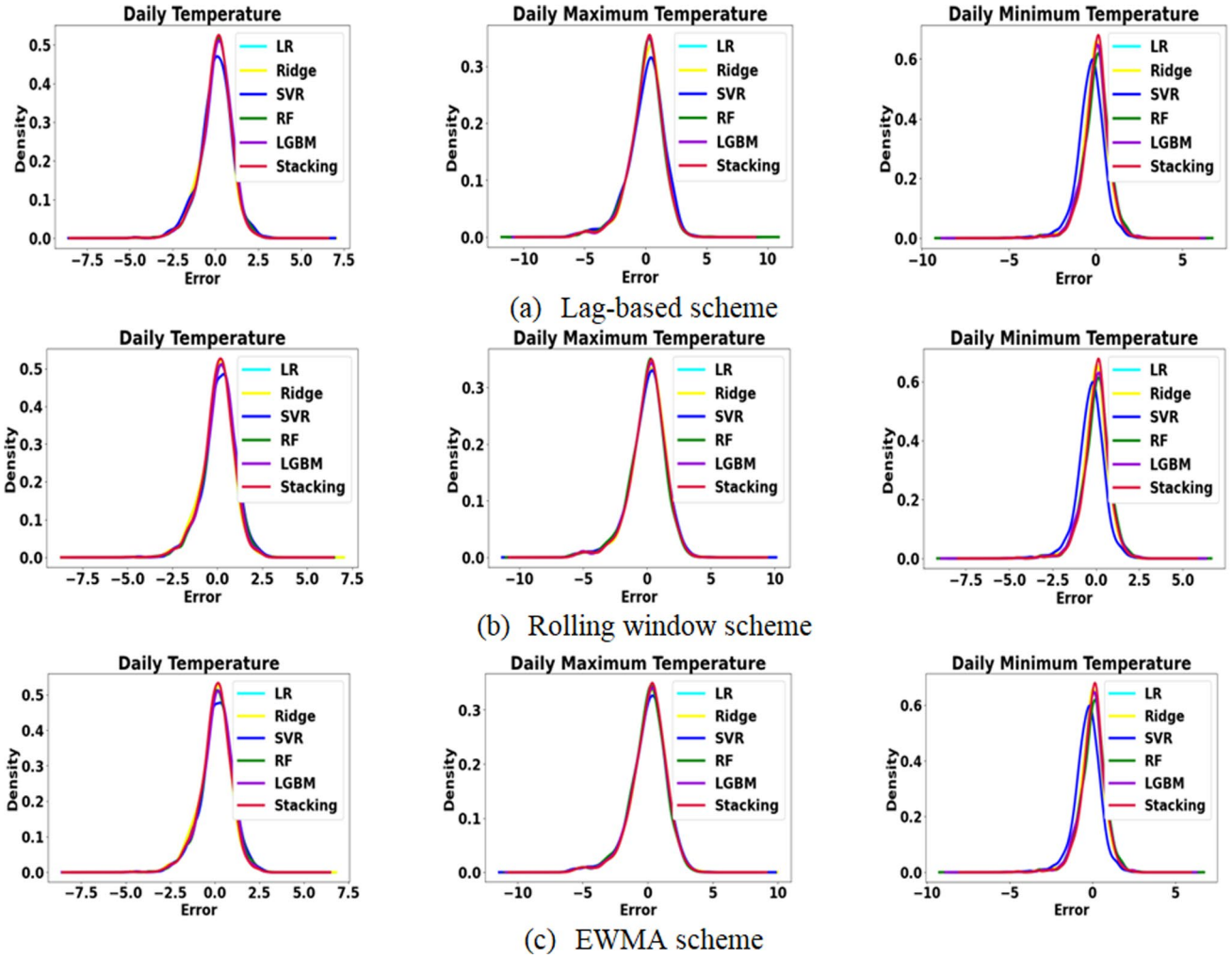

Fig. 8 Probability density functions showing the prediction error of the target variables obtained by the stacking model for three feature engineering statistical schemes: **a** lag-based, **b** rolling window, and **c** EWMA

Table 13 Standard deviation of predicted value of the target variables

Model	Lag-based scheme			Rolling window scheme			EWMA scheme		
	T	Tmax	Tmin	T	Tmax	Tmin	T	Tmax	Tmin
LR	0.895	1.380	0.748	0.893	1.380	0.747	0.890	1.380	0.747
Ridge	0.895	1.381	0.749	0.893	1.380	0.748	0.890	1.381	0.747
SVR	0.942	1.489	0.814	0.901	1.428	0.785	0.902	1.427	0.781
RF	0.880	1.391	0.786	0.881	1.373	0.784	0.883	1.376	0.785
LGBM	0.881	1.383	0.763	0.879	1.371	0.768	0.877	1.377	0.766
Stacking	**0.869**	**1.364**	**0.732**	**0.866**	**1.355**	**0.732**	**0.866**	**1.355**	**0.731**

The minimum standard deviations of the predicted values are presented in bold, which indicates a more accurate and reliable model performance

5.4 Trend analysis

Table 17 presents the outcomes of the Mann–Kendall (MK) and Sen's slope tests applied to the daily average, maximum, and minimum temperatures in 2021, based on both observed (BMD) and projected data. The BMD data reveals an upward trend (↑) in average temperatures from February to April (summer), with monthly increments ranging from 0.080 to 0.183 °C per day. Conversely, a downward trend (↓) is observed from November to January (winter),

Table 14 Time (second) for estimation of the models (the least estimation time is boldfaced)

Variable	Algorithm	Lag-based scheme	Rolling window scheme	EWMA scheme
Average temperature	LR	0.056	0.023	0.043
	Ridge	**0.007**	**0.014**	**0.015**
	SVR	3.939	1.007	0.978
	RF	42.295	7.857	6.633
	LGBM	0.928	0.114	0.068
	Stacking	79.362	48.193	33.928
Maximum temperature	LR	0.020	0.039	0.029
	Ridge	**0.012**	**0.009**	**0.011**
	SVR	2.054	1.151	1.006
	RF	9.337	4.149	1.242
	LGBM	0.180	0.089	0.686
	Stacking	56.560	36.986	17.864
Minimum temperature	LR	0.049	0.341	0.052
	Ridge	**0.027**	**0.038**	**0.015**
	SVR	8.102	7.699	8.353
	RF	8.128	3.690	11.129
	LGBM	3.899	4.970	1.176
	Stacking	63.698	40.858	70.409

Table 15 Results of Friedman tests

Weather variable	Critical value	p value	F_F value	Null hypothesis
Temperature	19.675	9.843e−07	48.904	Rejected
Maximum temperature	19.675	1.251e−07	53.862	Rejected
Minimum temperature	19.675	5.310e−07	50.397	Rejected

Table 16 Rankings of the algorithms using the Friedman test

Algorithm	Average rank		
	Average temperature	Maximum temperature	Minimum temperature
Stacking	1.500	1.000	1.000
LGBM	3.292	3.208	4.000
RF	3.375	2.625	5.000
LR	3.583	3.625	2.375
Ridge	3.750	4.542	2.625
SVR	5.500	6.000	6.000

accompanied by a reduction in Sen's slope ranging from − 0.200 to − 0.107 °C per day. No discernible trend (–) is noted from May to October (autumn). The predicted data, obtained by the stacking model using the EWMA scheme, is capable of mirroring these findings from the BMD data, though there are slight discrepancies, particularly in October. Notable trends in maximum average temperature are noted in February, March, April, and August, with temperature increases of 0.799, 0.200, 0.063, and 0.067 °C per day, respectively. Conversely, a decreasing trend was observed from December to January. The remaining months showed no substantial trends. The stacking model with a rolling window scheme yielded the most favorable outcomes for maximum temperature. The model effectively captured monthly trends comparable to those observed in the BMD data. Additionally, Sen's slope values are nearly identical in both cases.

Nevertheless, the minimum average temperature exhibited an upward trend from February to May, with temperature increments ranging from 0.048 to 0.272 °C per day. Likewise, a negative trend was observed, leading to a drop in the

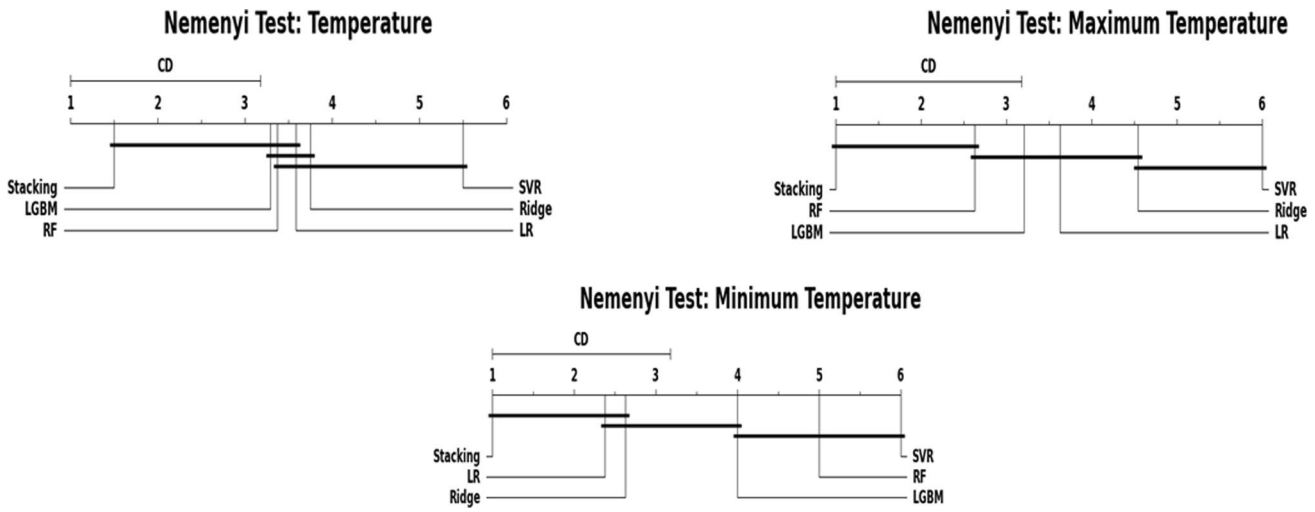

Fig. 9 CD diagram of the average performance rankings

Table 17 Mann–Kendall trend analysis results of daily temperature, maximum temperature, and minimum temperature for year 2021

Month	Tavg				Tmax				Tmin			
	Actual		Stacking		Actual		Stacking		Actual		Stacking	
	Trend	Slope	Trend	Slope	Trend	Slope	Trend	Slope	Trend	Slope	Trend	Slope
Jan	↓	−0.107	↓	−0.072	↓	−0.148	↓	−0.008	↓	−0.058	↓	−0.082
Feb	↑	0.183	↑	0.180	↑	0.799	↑	0.103	↑	0.272	↑	0.258
Mar	↑	0.160	↑	0.143	↑	0.200	↑	0.172	↑	0.160	↑	0.134
Apr	↑	0.080	↑	0.067	↑	0.063	↑	0.064	↑	0.048	↑	0.067
May	–	0.020	–	0.019	–	−0.030	–	−0.024	↑	0.062	↑	0.060
Jun	–	0.014	–	−0.003	–	−0.015	–	−0.007	–	0	–	0.009
Jul	–	−0.033	–	−0.013	–	−0.059	–	−0.034	–	0.007	–	0.013
Aug	–	0.013	–	0.026	↑	0.067	↑	0.062	–	−0.011	–	−0.005
Sep	–	−0.025	–	−0.019	–	−0.009	–	−0.021	–	−0.006	–	−0.018
Oct	–	−0.046	↓	−0.041	–	−0.007	–	−0.029	↓	−0.0	↓	−0.051
Nov	↓	−0.117	↓	−0.101	–	−0.010	–	−0.029	↓	−0	↓	−0.051
Dec	↓	−0.200	↓	−0.200	↓	−0.119	↓	−0.172	↓	−0.238	↓	−0.229
Annual	–	0.003	–	0.001	–	0.001	–	0.001	↑	0.0064	↑	0.006

minimum temperature from October to January. June to September did not display any notable trends. The projected data generated by the stacking model using the EWMA scheme mirrored the trends observed in the BMD data. Sen's slope values closely align in both cases. The annual trend analysis was conducted, and the findings indicate a rise in minimum temperature throughout the year 2021 in Cox's Bazar. However, no discernible trends were observed for both average and maximum temperatures.

6 Conclusion

Predicting the weather is a challenging and intricate endeavor. In this study, a perceptible stacking model was implemented to estimate daily air temperature parameters, which include average, maximum, and minimum temperature in Cox's Bazar. The results showed that the suggested stacking model performed better overall than the other optimal base models across all metrics, including MSE, MAE, MAPE, and R^2. In terms of forecasting T, the stacking model demonstrated MSE improvement rates of 13.2%, 8%, and 7.8% across the three schemes compared to SVR. Additionally, the MAE

improvement rates stood at 5.5%, 4.8%, and 4.7%. When it comes to forecasting Tmax, the MSE improvement rates ranged from 19.6 to 36%, and the MAE increment rates ranged from 4.9 to 11%. Similarly, for Tmin, the MSE improvement rates varied from 16.4 to 25.8%, and the MAE increment rates ranged from 7.7 to 12.1%. Among various statistical approaches, EWMA and rolling window schemes take the lead compared to the lag-based ones. Furthermore, statistical tests validate the stacking model's performance. Ultimately, the comprehensive annual trend analysis conducted provides invaluable insights into future trends, offering crucial information on whether they are set to ascend or remain stable. Therefore, the suggested stacking model is sufficient and a reasonable match for the data, which can be effectively integrated into web or mobile applications in various sectors, including agriculture and renewable energy planning.

While this study has shed light on the dynamics of temperature forecasting, it is important to acknowledge its limitations. The stacking model can be computationally expensive as it requires training multiple base models and a meta-model while dealing with large datasets (7,670 samples). Training the base models alongside the meta-model consumes more time compared to training a single model, which can be impractical for applications requiring real-time or time-sensitive responses. The model can be susceptible to overfitting, particularly in situations with limited sample sizes. Furthermore, the reliability of BMD meteorological data is compromised by instrument-related issues, data management errors, transparency deficits, and the potential for errors, inconsistencies, and data gaps, all of which undermine trust in its integrity. Addressing these constraints in future research promises to generate a more robust and reliable model for real-life applications. Future research could delve into more granular data sources, such as high-resolution satellite data, to address the spatial limitations we encountered. Additionally, to generate a temperature prediction challenge that involves forecasts over several time steps and encompasses various locations, we will carry out experiments at numerous diverse sites with the impact of temperature during various seasons. Advanced machine learning and deep learning methods can be adopted, which may help mitigate the uncertainties associated with temperature prediction. Collaborative efforts with meteorological experts could further enhance the accuracy of temperature forecasts.

Author contributions Conceptualization, TM; investigation, TM; methodology, TM; project administration, TM, GH, and SRS; software, TM; supervision, GH, and SRS; validation, TM, GH, and SRS; writing—original draft, TM; writing—review and editing, TMand SRS. All authors have read and agreed to the published version of the manuscript.

Funding This research received no external funding.

Data availability On request, we will provide the information.

Declarations

Ethics approval and consent to participate Not applicable.

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Alomar MK, et al. Data-driven models for atmospheric air temperature forecasting at a continental climate region. PLoS ONE. 2022;17(11): e0277079. https://doi.org/10.1371/journal.pone.0277079.
2. Lin M-L, Tsai CW, Chen C-K. Daily maximum temperature forecasting in changing climate using a hybrid of Multi-dimensional complementary ensemble empirical mode decomposition and radial basis function neural network. J Hydrol Reg Stud. 2021;38: 100923. https://doi.org/10.1016/j.ejrh.2021.100923.
3. Paul S, Roy S. Forecasting the average temperature rise in Bangladesh: a time series analysis. J Eng Sci. 2020;11(1):83–91. https://doi.org/10.3329/jes.v11i1.49549.

4. Roy M, Biswas B, Ghosh S. Trend analysis of climate change in Chittagong station in Bangladesh. Int Lett Nat Sci. 2015;47:42–53. https://doi.org/10.56431/p-7v90xn.
5. Apaydın M, Yumuş M, Değirmenci A, Karal Ö. Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data. Pamukkale Univ J Eng Sci. 2022;28(5):737–47. https://doi.org/10.5505/pajes.2022.66915.
6. Hanoon MS, et al. Developing machine learning algorithms for meteorological temperature and humidity forecasting at Terengganu state in Malaysia. Sci Rep. 2021;11(1):18935. https://doi.org/10.1038/s41598-021-96872-w.
7. Azamathulla HMd, Rathnayake U, Shatnawi A. Gene expression programming and artificial neural network to estimate atmospheric temperature in Tabuk, Saudi Arabia. Appl Water Sci. 2018;8(6):184. https://doi.org/10.1007/s13201-018-0831-6.
8. Lu M, et al. A stacking ensemble model of various machine learning models for daily runoff forecasting. Water. 2023;15(7):1265. https://doi.org/10.3390/w15071265.
9. Karevan Z, Suykens JAK. Spatio-temporal stacked LSTM for temperature prediction in weather forecasting. arXiv; 2018. http://arxiv.org/abs/1811.06341. Accessed 05 June 2023.
10. Roy DS. Forecasting the air temperature at a weather station using deep neural networks. Procedia Comput Sci. 2020;178:38–46. https://doi.org/10.1016/j.procs.2020.11.005.
11. Lee S, Lee Y-S, Son Y. Forecasting daily temperatures with different time interval data using deep neural networks. Appl Sci. 2020;10(5):1609. https://doi.org/10.3390/app10051609.
12. Mohammadi B, Mehdizadeh S, Ahmadi F, Lien NTT, Linh NTT, Pham QB. Developing hybrid time series and artificial intelligence models for estimating air temperatures. Stoch Environ Res Risk Assess. 2021;35(6):1189–204. https://doi.org/10.1007/s00477-020-01898-7.
13. Zhou J, Wang D, Band SS, Mirzania E, Roshni T. Atmosphere air temperature forecasting using the honey badger optimization algorithm: on the warmest and coldest areas of the world. Eng Appl Comput Fluid Mech. 2023;17(1):2174189. https://doi.org/10.1080/19942060.2023.2174189.
14. Nketiah EA, Chenlong L, Yingchuan J, Aram SA. Recurrent neural network modeling of multivariate time series and its application in temperature forecasting. PLoS ONE. 2023;18(5): e0285713. https://doi.org/10.1371/journal.pone.0285713.
15. Ke N, Shi G, Zhou Y. Stacking model for optimizing subjective well-being predictions based on the CGSS database. Sustainability. 2021;13(21):11833. https://doi.org/10.3390/su132111833.
16. Gu J, Liu S, Zhou Z, Chalov SR, Zhuang Q. A stacking ensemble learning model for monthly rainfall prediction in the Taihu Basin, China. Water. 2022;14(3):492. https://doi.org/10.3390/w14030492.
17. Zhu X, Hu J, Xiao T, Huang S, Wen Y, Shang D. An interpretable stacking ensemble learning framework based on multi-dimensional data for real-time prediction of drug concentration: the example of olanzapine. Front Pharmacol. 2022;13: 975855. https://doi.org/10.3389/fphar.2022.975855.
18. Salah S, Alsamamra HR, Shoqeir JH. Exploring wind speed for energy considerations in Eastern Jerusalem-Palestine using machine-learning algorithms. Energies. 2022;15(7):2602. https://doi.org/10.3390/en15072602.
19. Shabbir M, Chand S, Iqbal F. Bagging-based ridge estimators for a linear regression model with non-normal and heteroscedastic errors. Commun Stat Simul Comput. 2022. https://doi.org/10.1080/03610918.2022.2109675.
20. Alkhayat G, Mehmood R. A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy AI. 2021;4: 100060. https://doi.org/10.1016/j.egyai.2021.100060.
21. Perera A, Mudannayake SD, Azamathulla H, Rathnayake U. Recent climatic trends in Trinidad and Tobago, West Indies. Asia Pac J Sci Technol. 2020;25(2):1–11.
22. Erdebilli B, Devrim-İçtenbaş B. Ensemble voting regression based on machine learning for predicting medical waste: a case from Turkey. Mathematics. 2022;10(14):2466. https://doi.org/10.3390/math10142466.
23. Surakhi O, et al. Time-lag selection for time-series forecasting using neural network and heuristic algorithm. Electronics. 2021;10(20):2518. https://doi.org/10.3390/electronics10202518.
24. Cao XH, Stojkovic I, Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinform. 2016;17(1):359. https://doi.org/10.1186/s12859-016-1236-x.
25. Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. arXiv; 2019. http://arxiv.org/abs/1912.06059. Accessed 08 Sept 2023.
26. Huang Y, Zhao H, Huang X. A prediction scheme for daily maximum and minimum temperature forecasts using recurrent neural network and rough set. IOP Conf Ser Earth Environ Sci. 2019;237: 022005. https://doi.org/10.1088/1755-1315/237/2/022005.