Journal of Air Pollution and Health (Summer 2023); 8(3): 381-398

Original Article

Available online at http://japh.tums.ac.ir

Ensemble learning models for the prediction of the weekly peak of $PM_{2.5}$ concentration in Algiers, Algeria

Sabri Ghazi[1,], Ahmed Dib[2], Mohamed Said Mehdi Mendjel[1], Tarek Khadir[1], Julie Dugdale[3]*

[1] Electronic Document Management Laboratory (LabGED), Department of Computer Science, University Badji Mokhtar, Annaba, Algeria
[2] System and Networking Laboratory (LRS), Department of Computer Science, University Badji Mokhtar, Annaba, Algeria
[3] University Grenoble Alpes, Grenoble Informatics Laboratory (LIG), France

ARTICLE INFORMATION

Article Chronology:
Received 12 May 2023
Revised 23 July 2023
Accepted 01 September 2023
Published 29 September 2023

Keywords:
Particulate matters ($PM_{2.5}$); Air pollution; Ensemble learning; Time series forecasting; Air pollution prediction

CORRESPONDING AUTHOR:

sabri.ghazi@univ-annaba.dz
Tel: (+213) 6 97458559
Fax: (+213) 6 97458559

ABSTRACT

Introduction: This paper focuses on the prediction of weekly peak levels of Particulate Matter with an aerodynamic diameter of less than 2.5 µm ($PM_{2.5}$), using various Machine Learning (ML) models. The study compares ML models to deep learning models and emphasizes the explain ability of ML models for $PM_{2.5}$ prediction.

Materials and methods: We examine different combinations of features and time window dimensions to evaluate the performance of ML models. It utilizes Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Decision Tree (DT), and five Ensemble Models (EL) including AdaBoost, XGBoost, LightGBM, CatBoost, and Random Forest (RF). The dataset includes three years of daily measurements of weather parameters and $PM_{2.5}$.

Results: Lagged values of $PM_{2.5}$ improves prediction performance, particularly when the lagged value window size spans seven days or multiples thereof. This confirms that road traffic, which exhibits a weekly seasonality, is the primary source of $PM_{2.5}$ in Algiers. Interestingly, including lagged values of weather parameters decreases prediction performance, even when chosen based on their correlation with $PM_{2.5}$. The AdaBoost model performs the best, achieving a Root Mean Squared Error (RMSE) of 2.899 µg/m³ and an R^2 value of 0.96.

Conclusion: EL models, specifically AdaBoost, exhibit strong performance in predicting $PM_{2.5}$ levels. They not only provide accurate predictions but also allow analysis of feature importance. Lagged values of $PM_{2.5}$ have a greater impact on predictions compared to weather parameters. Surprisingly, including weather parameters hampers prediction performance. Therefore, the utilization of ensemble learning models offers valuable insights into feature significance.

Introduction

The degradation in air quality is a major challenge facing many cities in the world. In developing countries, uncontrolled urban expansion, fossil energy-based transportation, and the lack of legislation to enforce air quality standards,

Please cite this article as: Ghazi S, Dib A, Mendjel MSM, Khadir T, Dugdale J. Ensemble learning models for the prediction of the weekly peak of $PM_{2.5}$ concentration in Algiers, Algeria. Journal of Air Pollution and Health. 2023;8(3): 381-398.

lead to alarming levels of air pollution. Peak periods occur when the concentrations of air pollutants are above the tolerated level. Among the pollutants responsible for these peaks is the Particulate Matters ($PM_{2.5}$), which is a mixture of solid and liquid substances, mainly generated by anthropogenic activities. The combustion engine, construction, industrial process, and agriculture are among the main sources of $PM_{2.5}$. In Algiers Fe and Sc are highly present among the heavy metal content of $PM_{2.5}$ as concluded in research [1]. The authors deduced that the annual level of $PM_{2.5}$ exceeds local and international standards. The same conclusion was reached in [2] where the authors analysed samples of PM_1, $PM_{2.5}$, and PM_{10} from two stations in Algiers during 2015 and 2016 in an urban and roadside environment. By analyzing the samples of $PM_{2.5}$, the concentrations of heavy metals were determined, with Pb representing 5%. The composition of PM_{10} and $PM_{2.5}$ in an urban area in Algiers is also described in [3]. The heavy metal content of $PM_{2.5}$ confirmed that the origin was from road traffic and Saharan dust. It is worth mentioning that Algeria was the only country in the world that continued to use leaded carburant until August 2021 when the Algerian government passed a law banning the use and sale of leaded carburant. Due to its diameter and toxicity, $PM_{2.5}$ can be inhaled by humans leading to serious health problems [4]. Therefore, having an accurate prediction of the peak periods of $PM_{2.5}$ can help decision-makers mitigating the crisis and reducing its effects, specifically by warning people who have special medical conditions.

Ensemble Learning (EL) models are easy to implement, require less computation, and are explainable when compared to Deep Learning models. We chose to use EL models because of their transparency in terms of feature importance. These models often provide explicit feature weights or coefficients that indicate the contribution of each feature in predicting the target variable.

This paper has a specific focus on investigating Ensemble Learning (EL) models and their performance in predicting the weekly peak of $PM_{2.5}$. The primary aim of this study is to thoroughly evaluate the effectiveness of these EL models in accurately forecasting the highest concentration levels of $PM_{2.5}$ over a weekly timeframe. By doing so, we intend to provide valuable insights into the potential applications of ensemble learning for this particular environmental forecasting task. In addition to assessing the predictive capabilities of EL models, we also seek to delve into their inner workings and enhance our understanding of their functionality. This secondary objective involves a detailed examination of the importance of individual features and their lagged values within the ensemble learning models. By analysing the contribution of each feature and their lagged values, we aim to gain a deeper understanding of the factors that significantly influence $PM_{2.5}$ levels during the weekly peak periods, and the time dependency that may exist between a lagged value of a feature and the $PM_{2.5}$ level. This exploration of feature importance will aid in interpreting the model's outputs more effectively and comprehensively. The ultimate goal of this work is to enhance the interpretability of the ensemble learning models, making the predictions more accessible and useful for various stakeholders and policymakers. Having a clear understanding of the underlying factors and the reasoning behind the model's forecasts is crucial for making well-informed decisions related to $PM_{2.5}$ levels. By achieving a higher level of interpretability, this study aims to contribute to the field of environmental forecasting and support decision-making processes that can positively impact air quality management and public health.

Table 1. Studies presenting PM2.5 concentration prediction

Area and Period	Prediction horizon	Model	Features Engineering and hyperparameters	Lagged values	Inputs	Multi /single output
Algiers, 2003-2004	24 h	MLP	Correlation	-	$PM_{2.5}$,WS, RH, T	single
Algiers, 4 months	Not mentioned	SVM	Correlation, Dragonfly	-	$PM_{2.5}$,WS, RH, T, P	Single
Taichung, Taiwan, 2017	3 h	AIF	Hierarchical Clustering	-	WS, RH,T, P, A	Single
Gansu, China, 2019-2020	From 1 h to 48 h	G-LSTM	Adjacency Matrix	4 h	$PM_{2.5}$, WS, WD, RH, T, P, Pr, CO,NO_2, O_3,SO_2 PM_{10}, $PM_{2.5}$	Both
Wayne, Michigan, USA	-	Lag-FLSTM	Bayesian optimization	48 h	PM2.5, WS, WD, Press, T, CO, SO_2, NO_2, PM_{10}	Single
Beijing, China	From 24 h to 168 h	AE-Bi-LSTM	Auto-Encoder	-	$PM_{2.5}$, WS, P, Snow, T, Dewpoint	Single
Beijing, China, 3 years 2015-2017	-	CNN-LSTM	Mutual Information estimator	-	CO, SO_2, O_3, NO2, $PM_{2.5}$, PM10, T, WD, WS,	Single
Beijing-Tianjin-Hebe, China	From 1 h to 24 h	LSTM	Regression Tree, ANN	-	WS, WD, RH, T, Workday/Weekend, Pres,P,, Dew point, Season, Month, $PM_{2.5}$	Single
2015-2016	-	OrdinaryDifferential Equation	Genetic Algorithm	-	$PM_{2.5}$	Single
London, UK, 2004-2013	1 h	Linear Regression, Random Forest	Generalized Additive Model	-	NO_2, PM_{10}, $PM_{2.5}$, Latitude, T, Week day, WS, WD, RH, Roadside vs Background	Single
France, 2000-2019	1 day	Gaussian Markov Random Field,Random Forest,	Generalized Additive Model	-	PM_{10}, $PM_{2.5}$, AOD, P, T, WS.	Single
Italy, 2013-2015	1 day	Random Forest	-	-	AOD, $PM_{2.5}$, PM_{10}, WD, WS, Press, P, T	Single
Iran, Tehran, 2015-2018	1 day	Random Forest, XGBoost	RF Features Importance XgBoost Features Importance Permutation Importance.	2 Days	AOD, $PM_{2.5}$, WS, RH,WD, P, Press, T, Dew Point	Single
Beijing, China, 2018	1 day	Temperature-Nased Deep Belief Networks	-	-	WS, P, T, PM_{10}, SO_2, CO_2, Pess, RH	Single
Newport, Taiwan, 2012-2017	-	XGBoost, RF, MLP, Decsion Tree, K neares neighbours	-	-	-	Single
Wrocław, Poland, 2015-2016	1 hour	RF	RF Features Importance.	-	Road Traffic, T, WS, WD, RH, Press, week day, holidays, month.	Single
Christchurch, New Zealand	1 h peak 1 day peak	boosted gradient machine	-	-	T, WS,NO, NO_2	Single
Hohhot, Harbin, Wuhan, Changsha China	1 h , 2 h , and 3 h	outlier robust extreme learning machine	-nonconvex sparse regularization -wavelet transform	-	-	Multi

Related works

Many approaches have been used to predict $PM_{2.5}$ concentrations and they can be clustered into five categories: Deterministic models, Linear models, Machine learning based models, Hybrid models, and Satellite-derived Aerosol Optical Depth models [5]. Moreover, they can be categorised according to: the model inputs; the prediction horizon and the studied region. A non-exhaustive review of recent studies proposing models to predict $PM_{2.5}$ concentration is summarized in Table 1. In the table, weather parameters such as Wind Speed, Wind Direction, Relative Humidity, Pressure, Ambient temperature and Cumulative precipitation are noted respectively as WS, WD, RH, Pr, T and P. Anthropogenic event data is noted as A. Despite its significant impact on the city's air quality and a lack of measurements on $PM_{2.5}$ concentration in Algiers, there are a limited number of studies that present models to forecast $PM_{2.5}$ in Algiers. An MLP (Multi-Layer Perceptron) model to predict the long-term concentration of PM_{10} in Algiers is proposed in [6]. It was trained using a two year dataset of PM_{10} concentration and meteorological parameters (wind speed, relative humidity, and temperature), that were selected based on their correlation with PM_{10}. However, the dataset is relatively old (2003-2004) and does not reflect the climatic and the anthological changes that have occurred during the recent decades in Algiers. An (Support Vector Machine) SVM model to predict the concentration of PM of different sizes, including $PM_{2.5}$, in Algiers is described in [7]. To select the best model hyper-parameters, the authors used a swarm algorithm called Dragonfly. The model showed relatively good performances. However, the dataset is limited as it only covers four months and does not include the yearly seasonality of $PM_{2.5}$. Therefore, it affects the model generalization. An ordinary differential equation to model $PM_{2.5}$ is studied in [8]. The authors compared it with an autoregressive model and showed a relatively similar performance. However, the model was trained using a limited dataset covering only two months of daily $PM_{2.5}$ concentration, leading to a weak generalization. Machine learning models are commonly used and are compared with linear models. To eliminate short-term fluctuations that affect the accuracy of the prediction. The $PM_{2.5}$ times series is smoothed using wavelet transformation [9]. To mitigate the effects of the sudden change in climatic parameters and anthropogenic events, the authors described in [10], the use of an unsupervised method to cluster anthropogenic and environmental events. They found that unexpected events such as rainfall intensity, as well as wind speed, and road traffic have an impact on the concentration of $PM_{2.5}$. Event data are collected from the forecast error of an Adaptive Iterative Forecast model. To tackle the lack of $PM_{2.5}$ measurement in London, the authors in [11] developed a $PM_{2.5}$ concentration prediction model. The model uses the concentration of PM_{10} and NO as inputs. Linear regression and Random Forest models are combined using GAM (Generalized Additive Model). The authors combined weather parameters to obtain the best-performing model. To predict $PM_{2.5}$ in Beijing, China, authors in [12], used TDBN (Temperature-based Deep Belief Networks) with multiple hidden layers and various sizes. The inclusion of topographical data as input to forecast $PM_{2.5}$ in Newport, Taiwan is presented in [13]. The performance of RF in predicting of $PM_{2.5}$ was investigated in [14]. The dataset was divided into many subsets, and after assessing the accuracies of each one, the authors concluded that RF is more accurate at predicting $PM_{2.5}$ during warmer periods. A binary classification approach is used to predict $PM_{2.5}$ excess in [15]. The $PM_{2.5}$ measurements were converted into two classes: Peak and No-Peak. However, since the number of peaks is always lower than the normal level, this results in an unbalanced dataset as the peaks represent a minority class. This, in trun, affects the model's generalizability. Recent studies have used deep learning models of various architectures to prediction $PM_{2.5}$. Many monitoring stations in Gansu, China, are modelled as weighted graphs, and a Long Short-

Term Memory (LSTM) network is designed for each station. The weight on the edge between two stations is included in the LSTM input of another station. The model can forecast $PM_{2.5}$ concentration in every station without the need to build a model for each station. According to the study [16], the model took into consideration the spatiotemporal information and, as a result, outperformed the ensemble learning model, using the same dataset. A Bayesian optimization was used to determine the values of the hyper-parameters of a fully connected LSTM model. The model used lagged values of inputs, including the weather parameters. Compared with other models using the same dataset, the model gave the best performance. However, the used data to validate the model was randomly selected as described in [17]. With time series data, this could lead to poorly explanatory data since it lacks the time order of each observation. The authors in [18] used an Auto-Encoder to compress the feature space before passing it as input to an LSTM model. The later receives as input the lagged values of $PM_{2.5}$, snow, precipitation, ambient temperature, wind speed, and direction. Compared to classic models such as CAMx, CMAQ, and other deep learning models, the proposed model demonstrated the best performance. The authors argue that for a long-term prediction, the model trained using only $PM_{2.5}$ performed better than the one that includes weather parameters. However, for a short prediction horizon, the models that included weather parameters exhibited better precision. A Mutual Information Estimator for determining the correlation between times series of weather pollutant parameters from 384 stations across China is presented in [19]. The authors claim that this helps to capture spatiotemporal information. The selected features were then used to train a CNN-LSTM model. A multi-stage method to consider spatial and temporal information in the prediction of $PM_{2.5}$ is described in [20]. Initially, for each monitoring station using LSTM, a spatial predictor and a temporal predictor are trained. Secondly, the output of each LSTM model is used in a Regression Tree model to predict $PM_{2.5}$ concentration. Lastly, an ANN (Artificial Neural Network) is used to predict $PM_{2.5}$ concentration at a grid level. Some studies included additional inputs such as AOD (Aerosol Optical Depth). As reported in [21] AOD and empirical data are used to predict the daily PM_{10} and $PM_{2.5}$ concentrations in France. A RF model is used to impute the concentration of $PM_{2.5}$ in the stations that measure only PM_{10} concentration. The missing values of AOD are also predicted using an RF model. GAM is used to combine the outputs of Linear Regression, RF and GMRF (Gaussian Markov Random Field). The same strategy is used in [22] to predict PM_{10} and $PM_{2.5}$ in Italy, with the adding of a local predictor at the last stage to improve the prediction at a short time horizon. Using data from Tehran, Iran, authors in [23], investigated the contribution of AOD to enhance the performance of $PM_{2.5}$ prediction model.

Background on ensemble learning approach

A Decision Tree (DT) is a machine-learning model that builds a tree by inducing the rules from the data. First, it selects the feature that splits the training sample and builds a decision node, and recursively builds sub-trees. Feature selection is performed using the Gini impurity metric, which calculates how well a feature splits the samples. A DT is commonly used in ensemble learning, in which many models, called weak learners, are trained and their outputs are combined to obtain the final decision. Many techniques are used to combine the outputs. Bootstrap aggregation assigns an equal weight to each model output in the vote to determine the final output. A Random Forest (RF) [24] uses DT models and combines their outputs. Many DT models are trained using random samples of the training data and random subsets of the features. The AdaBoost algorithm, short for adaptive boosting [25] determines the parameters by re-assigning the weights to each instance, with higher weights to incorrectly classified instances. XGBoost [26] is DT-based model that uses a gradient boosting strategy. It applies the principle of boosting and provides a parallel tree boosting. LightGBM is

a recent improvement of the gradient boosting algorithm [27]. Its principal advantage over the other gradient boosting algorithms is its ability to resolve the scalability problem by adopting a leaf-wise tree growth strategy. It splits the tree leaf-wise with the best fit whereas other boosting algorithms split the tree depth-wise or level-wise. Therefore, when growing on the same leaf in LightGBM, the leaf-wise algorithm can reduce loss more than the level-wise algorithm and hence results in much better accuracy, which is rarely achieved by any of the existing boosting algorithms. Another version of gradient descent is CatBoostGBM [28], which is a gradient descent algorithm designed to deal with categorical features and also avoid the overfitting problem.

Materials and methods

Studied region

Algiers is located in the centre of the north of Algeria, it is a coastal city bordered by the Mediterranean Sea on the North. It is the economic and the political capital of Algeria. According to the Office National of Statistics [29], in 2019 the estimated population was 8 million habitants. The city has a high economic attraction; it hosts many central administrations, international corporations' headquarters, and four active industrial zones. However, public transportation in Algiers has not expanded proportionally to the population growth, resulting in a heavy reliance on personal cars, according to a study by researchers, Algiers's motor fleet reached 2 million in 2019 . The city also has a seaport where goods are primarily transported using trucks [30].

Dataset description

This study uses a dataset covering 3 years (from 2019 to 2021) of daily measures of climatic parameters and $PM_{2.5}$ concentration. The measures of $PM_{2.5}$ were collected by EPA US-Embassy station in Algiers that has the following GPS coordinates 36.75595300548415; 3.039189599146588, the data is publiclly available from [31]. The climatic parameters are provided by the official Algerian meteorology agency (ONM). Table 2 describes some statistical properties. Some important events occurred during the period of the dataset, the first is the COVID-19 lockdown, which started in March 2020 to December 2020, and also during the second peak during August 2021. Moreover, the forest fires in the Tizi-Ouzou Mountains lasted for 7 days, from 9 to 15 August 2021. Fig. 1 illustrates a positive and negative correlation between $PM_{2.5}$ and the climatic parameters.

Auto-correlation

Fig. 2 shows the auto-correlation of $PM_{2.5}$, which measures the correlation between the lagged values and the current value of the $PM_{2.5}$ time series. The lagged values vary from 1 day to 35 days. As illustrated, the local peaks show a positive correlation between the value of the $PM_{2.5}$ and its past values, specifically the day numbers are multiples of seven such as 7, 14, 21, 28. This is also confirmed in Fig. 3, in which the weekly seasonality is clearly shown. This shows how $PM_{2.5}$ concentration decreases during Friday and Saturday, which is the Algerian local weekend. During the weekday the $PM_{2.5}$ concentration increases specifically, on Sunday, Monday, Tuesday and Thursday. With a local peak on Tuesday.

Table 2. Statistical properties of the dataset

	Mean	std	min	max	Missing value
$PM_{2.5}$	67.78	15.15	40,00	172.00	10%
Max_temperature (C°)	23.24	5.98	10,00	41.00	0%
Min_ temperature (in C°)	19.42	5.91	0,00	34.00	0%
Wind speed_Max_kmh	16.46	6.75	4,00	44.00	0%
Ttemperature _Morning (in C°)	18.69	5.67	0,00	33.00	0%
Temperature _Noon (in C°)	22.58	6.07	0,00	38.00	0%
Temperature _Evening (in C°)	21.53	5.97	0,00	40.00	0%
PreCIP_Total_Day (mm)	1.79	4.35	0,00	35.00	0%
Humidity_Max_ (%)	63.46	12.78	34,00	94.00	0%
Visibility_Avg (km)	9.90	1.05	6.875	20.00	0%
Pressure_Max_ (mega bar)	1018.69	5.30	1006.00	1035.00	0%
Cloud cover_Avg_(%)	28.80	25.28	0.00	94.375	0%
Heat index_Max (C°)	24.13	6.66	10.00	44.00	0%
Dew point_Max (C°)	14.46	4.79	2.00	26.00	0%
Wind temp_Max (C°)	19.20	6.28	4.00	34.00	0%
Weather_code_Morning	141.24	61.60	113.00	386.00	0%
Weather_code-Noon	140.42	61.32	113.00	386.00	0%
Weather-code-Evening	144.96	67.18	113.00	389.00	0%
Total_Snow_mm	0.00	0.00	0.00	0.00	0%
UV_Index	3.70	2.39	1.00	9.00	0%
Sun hour	10.48	3.02	3.5	14.5	0%

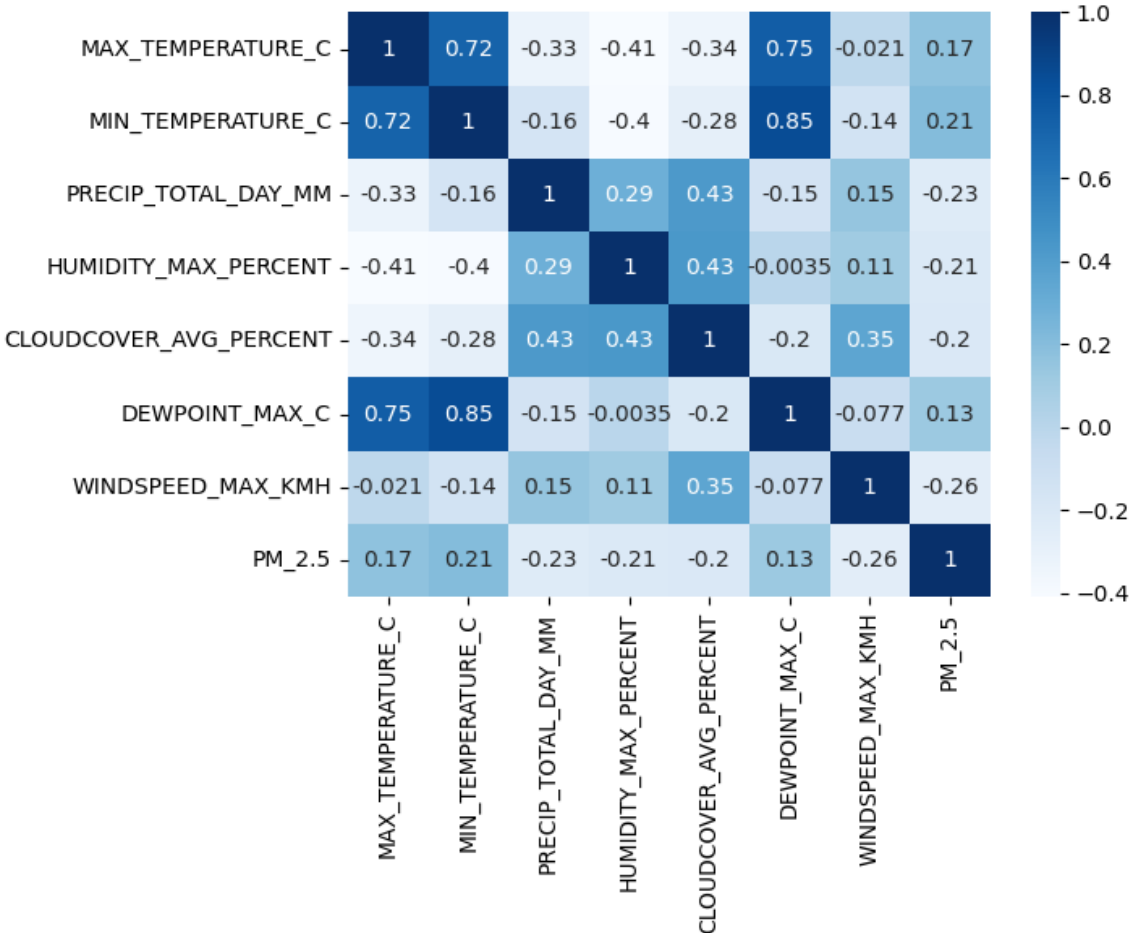

Fig. 1. Correlation between the features of the dataset

Data normalisation and missed values imputation

As described in Table 2, $PM_{2.5}$ times series contains 10% missed values. To maintain the time order and its impact, we imputed them using KNN (K Nearest Neighbours) imputer. This algorithm [32] uses a Euclidian distance to determine the K closest complete samples of the dataset. Then it fills in the missed values with a weighted average of the neighbours. Since the features are in different scales, we normalized the data using Eq. 1.

$$X_{scaled} = \frac{X - min\ (X)}{max(X) - min\ (X)} \quad (1)$$

Where min and max are functions which compute the minimum and maximum value, and X is the vector to be scaled.

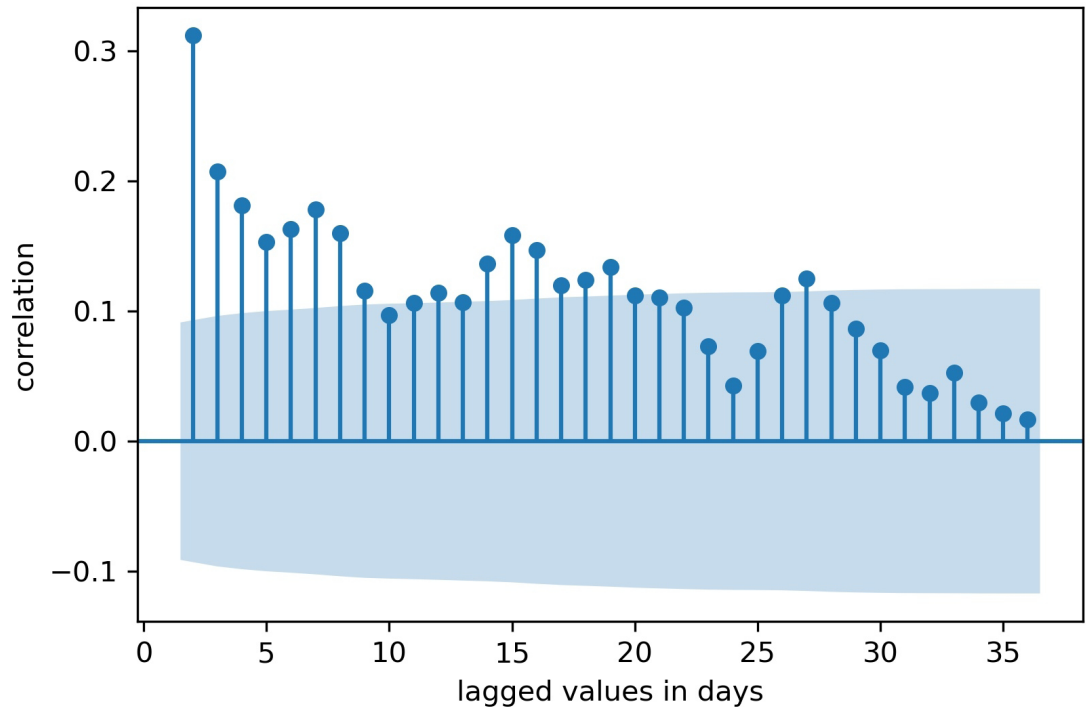

Fig. 2. Auto-correlation of $PM_{2.5}$ time series

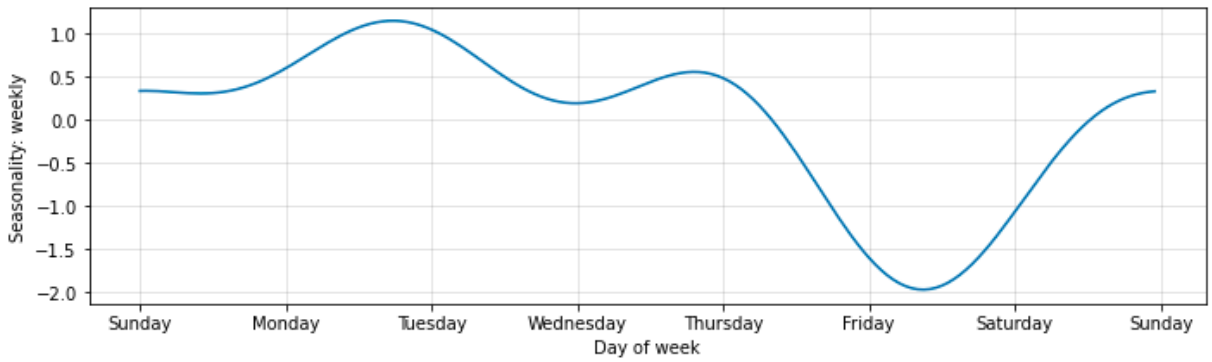

Fig. 3. Weekly seasonality of $PM_{2.5}$ in Algiers

Performances metrics

In order to compare the performance of the models, we used RMSE (Root of Mean Square Error), MAE (Mean Absolute Error), and R^2 (the coefficient of determination), as defined in (2) , (3), (4), respectively.

$$rmse = \sqrt{\frac{\sum_i^n \left(y_{i,measured} - y_{i,predicted}\right)^2}{n}} \quad (2)$$

$$MAE = \frac{\sum_i^n \left|y_{i,predicted} - y_{i,measured}\right|}{n} \quad (3)$$

$$R^2 = 1 - \frac{\sum_i^n \left(y_{i,measured} - y_{i,predicted}\right)^2}{\sum_i^n \left(y_{i,measured} - \acute{y}_{measured}\right)^2} \quad (4)$$

Where $y_{i,\ measured}$ is the i^{th} measured value of a vector of n values, $y_{i,\ predicted}$ is the i^{th} predicted value of the vector of n values. $y'_{measuredis}$ the mean of the measured value.

Model hyper-parameter

Hyper-parameters are configuration settings that are not learned from the data, but rather specified by the designer before training. Training machine learning models involves finding the best set of hyper-parameters that optimize the model's performance. In this study we used grid search, which is commonly used technique that allows systematically searching through a predefined grid of parameter values and finding the best combination.

Results and discussion

The objective is to design a model which maps the input $PM_{2.5}(t), PM_{2.5}(t-1) \ldots PM_{2.5}(t-k)$, $WeatherFactor_1(t), WeatherFactor_1(t-1) \ldots$ $WeatherFactor_1(t-k), \ldots WeatherFactor_m(t), WeatherFactor_m(t-1), \ldots$ $WeatherFactor_m(t-k)$ to the output representing the peak of the next week: $\max(PM_{2.5}(t+1), PM_{2.5}(t+2), PM_{2.5}(t+3), PM_{2.5}(t+4), PM_{2.5}(t+5), PM_{2.5}(t+6), PM_{2.5}(t+7))$. Where t is the day, WeatherFactor represents a weather factor, m is the number of the used weather factors, k is the number of lagged values, and max is a function that returns the maximum values of $PM_{2.5}$. To train the models we used the first 70% of the dataset, the remaining 30% were used to test the performance of the models. We computed the peak of each week of the dataset to form the target variable.

To investigate the most impactful features, we used four combinations: (1) All the features and their lagged values, (2) Univariable, which includes $PM_{2.5}$ lagged values, (3) Lagged $PM_{2.5}$ and some highly correlated climatic features, and (4) Lagged values of $PM_{2.5}$ and selected climatic features without their lagged values. All the aforementioned combinations were tested using lagged values from one day to 30 days, employing 8 Ensemble Learning (EL) models: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Decision Tree (DT), and five Ensemble Models, including AdaBoost, XGBoost, LightGBM, CatBoost, and Random Forest (RF). The total number of trained and evaluated models is 960.

All the features and their lagged values

Lagged values of 21 climatic features, and $PM_{2.5}$ are used, from one day to 30 days. As described in Fig. 4, the best performing model is RF with RMSE of 3.648 and R^2 of 0.937, for lagged values of 7 days. The next best-performing model is AdaBoost with RMSE of 4.770 and R^2 of 0.892. The order changes with lagged values of 24 days, 28 days, 29 days, and 30 days, when LightGBM outperformed RF. For example, LightGBM shows RMSE of 4.566 and R^2 of 0.901 where RF is 4.832 and R^2 is 0.889 for the lagged value of 24 days. Fig. 4 shows the evolution of the RMSE according to the number of lagged values. Except for SVM, the performance of the other models starts to improve when inputs with a seven days lagged value are used. Fig. 5 shows the relative importance of the feature determined by using RF built-in method. The features X its j lagged value is noted $X_{t\_T\_j}$. $PM_{2.5}$ lagged values come first specifically $PM_{2.5}(t-1)$, $PM_{2.5}(t-5)$, $PM_{2.5}(t-2)$, $PM_{2.5}(t-3)$, $PM_{2.5}(t-4)$, after that comes the first climatic parameter Pressures_Max. As illustrated in Fig. 4, the lagged values of climatic parameters are not considered important in the RF model with a 7 day lagged value, which is the best performing model, except the dew-point and max temperature.

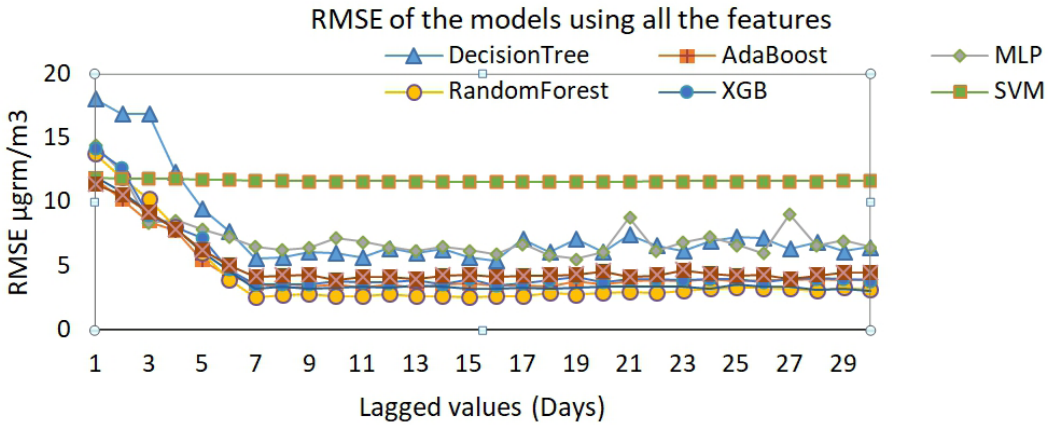

Fig. 4. The RMSE evolution according to the number of lagged values, models using all the features

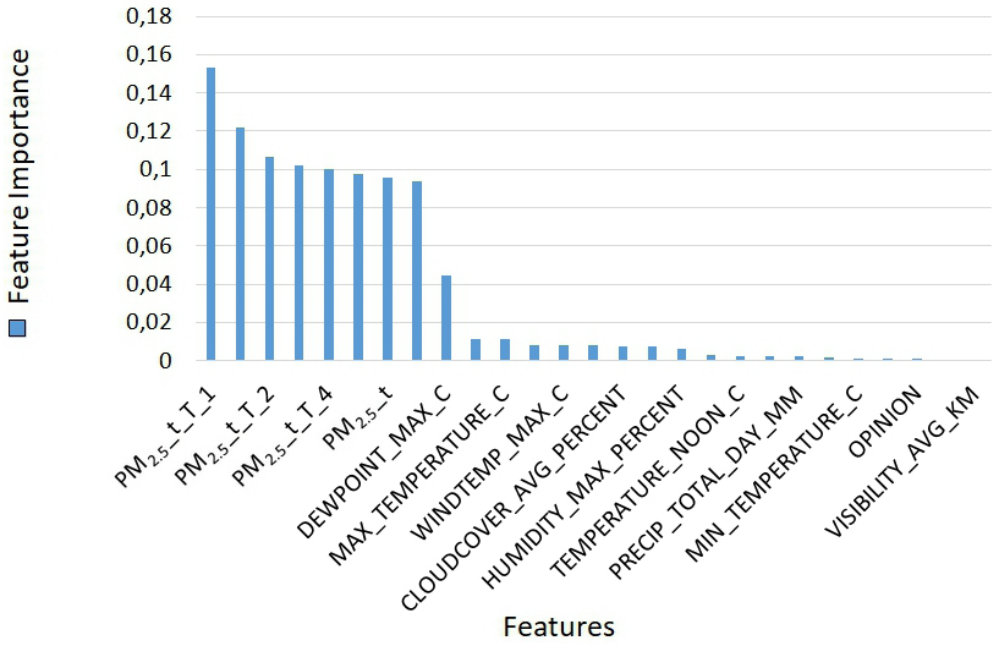

Fig. 5. The RF feature importance using a 7 day lagged value, all features included

Univariables models

The models receive the $PM_{2.5}$ lagged values as input. The size of the lagged values window was varied from 1 day to 30 days. For models with inputs of lagged values of 7 days, Adaboost showed the best performance with an RMSE of 2.899 and an R^2 of 0.960, followed by MLP with an RMSE of 2.915 and an R^2 of 0.959. RF exhibited an RMSE of 2.918 and an R^2 of 0.959. This implies that MLP and Adaboost perform best when using only the lagged values of the time series. The order changes for lagged values of 21, 22, 24, 25, 26, 27, 28, 29, and 30, in which LightGBM shows the best performance. For example, LightGBM with 25 days of lagged values demonstrated an RMSE of 3.791. The best model with 23 lagged values is RF, with an RMSE of 3.888. Fig. 6 displays the RMSE of the models trained using inputs with lagged values from 1 day to 30 days. Fig. 7 illustrates the feature importance of the RF model trained using 7 lagged values. The lagged values $PM_{2.5}$(t-1) and $PM_{2.5}$(t-5) seem to remain important.

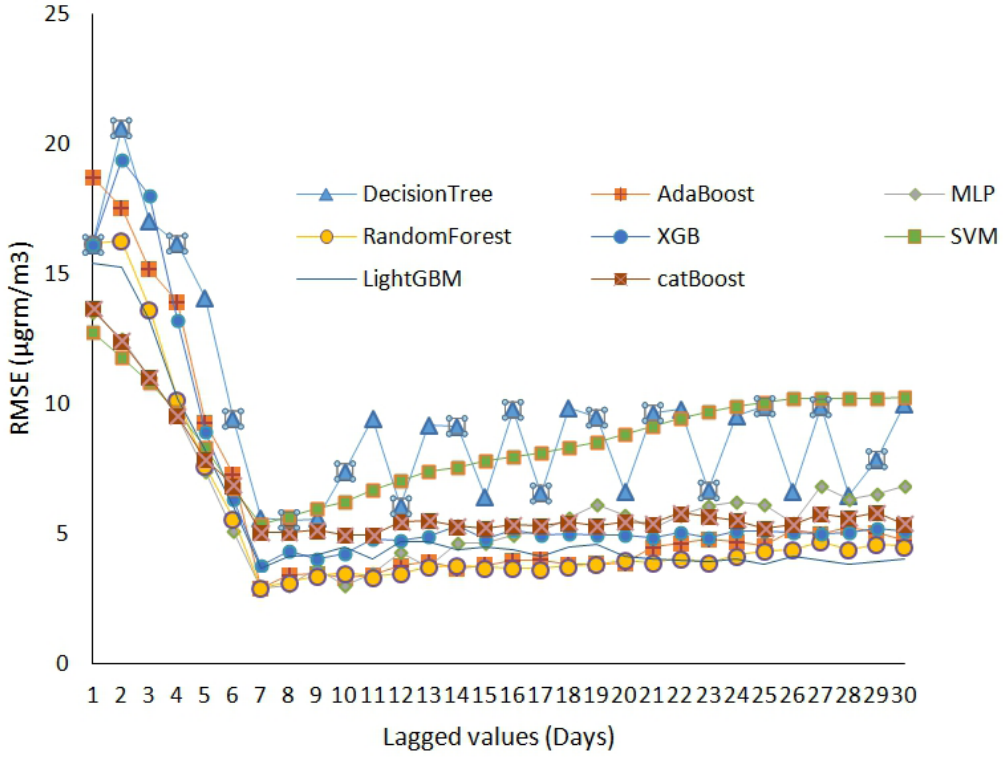

Fig. 6. The RMSE according to the number of lagged values of $PM_{2.5}$, models use only $PM_{2.5}$ no climatic parameters

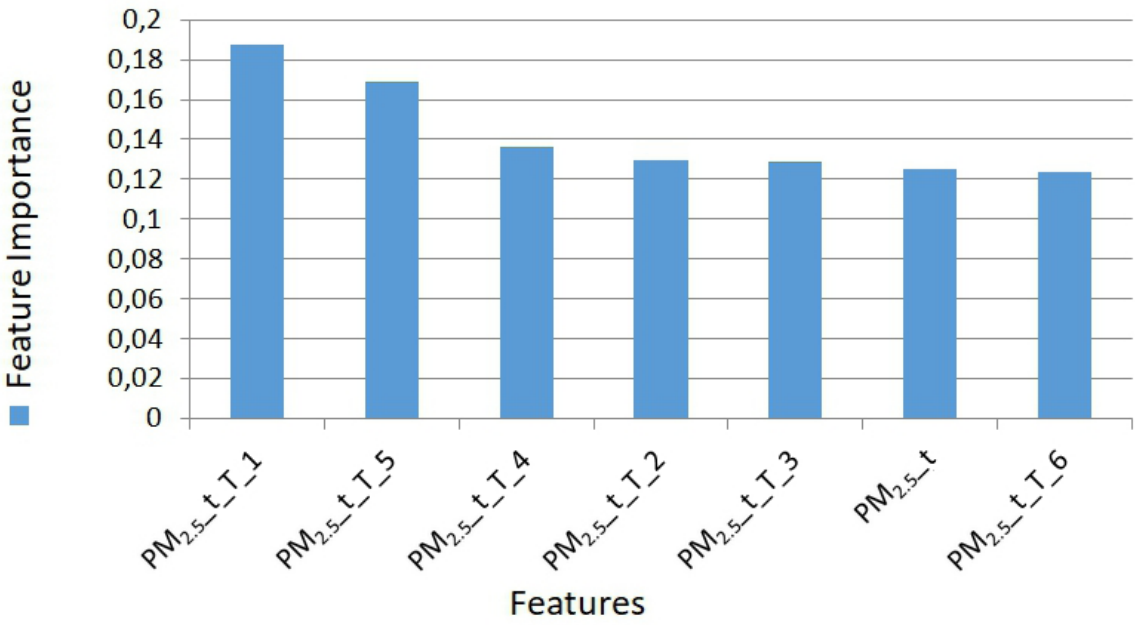

Fig. 7. Features importance of a RF model trained only with $PM_{2.5}$, with a 7 day lagged value

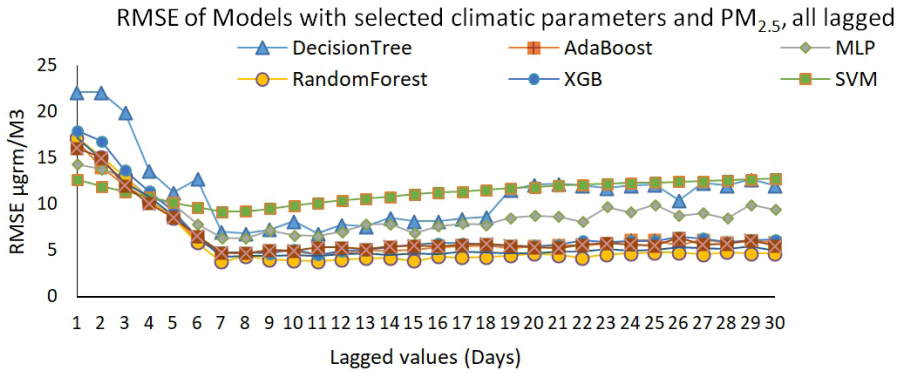

Fig. 8. RMSE evolution according to the size of the lagged values window

Table 3. Comparing the performance of our models with other research's models

Study	R^2	RMSE	MAE
This work - Selected features and ony lagged $PM_{2.5}$	0.956	3.039	2.495
This work - Lagged values of both selected features and $PM_{2.5}$	0.931	3.791	2.534
This work - Only $PM_{2.5}$	0.960	2.899	1.843
This work - All features included	0.937	3.648	2.551
Work of researchers (Ref 23)	0.8	9.93	13.58
Work of researchers (Ref 18), 168 hours ahead	-	7.93	-
Work of researchers (Ref 33)	-	3.58	7.44
Work of researchers (Ref 17)	-	3.482	1.85
Work of researchers (Ref 16)	-	3.405	2.60
Work of researchers (Ref 20)	0.87	24.24	8.25
Work of researchers (Ref 19)	-	2.870	2.11
Work of researchers (Ref 6)	0.85	13.780	-
Work of researchers (Ref 12)	0.86	11.190	12.29
Work of researchers (Ref 7)	0.98	1.926	-
Work of researchers (Ref 10)	-	1.780	1.300
Work of researchers (Ref 22)	0.81	5.36	-
Work of researchers (Ref 11)	0.95	-	-
Work of researchers (Ref 14)	0.57	-	-

Model using lagged $PM_{2.5}$ values and lagged values of selected features

The models were trained using inputs of $PM_{2.5}$ and selected climatic parameters, which were chosen based on their correlation with the target $PM_{2.5}$. Among all the tested combinations, the models using 7-day lagged values exhibited the best performance. The RF model demonstrated the best performance with an RMSE of 3.791 and an R^2 of 0.931, followed by LightGBM with an RMSE of 4.345. Fig. 8 depicts that the order changes with 8 lagged values, where LightGBM shows an RMSE of 4.3733, whereas RF shows an RMSE of 4.423. This order remains for lagged values of 23 days and 27 days. Fig. 9 displays the feature importance of the best-performing model. As illustrated, the $PM_{2.5}$ lagged values come first, specifically $PM_{2.5}$(t-5) and $PM_{2.5}$(t-1). After that, temperature appears to be the most significant weather factor.

Lagged values of $PM_{2.5}$ and selected climatic features without their lagged values

We trained the models using an input composed of lagged values of $PM_{2.5}$ and no lagged values of the selected weather parameters. This was done in order to determine how much the lagged values of weather factors can impact the models' performance. As shown in Fig. 10, the best performing model was found to be the MLP for 8 day lagged values, with a RMSE of 3.039, followed by RF with a RMSE of 3.505.

The order changes with 9 lagged values, the latter shows a RMSE of 3.386 and MLP shows 3.623. LightGBM outperforms both models for 22 and 27, 28 and 29 lagged values, for example with 29 lagged values it shows a RMSE of 3.934. Fig. 11 illustrates the importance of the features, it is noted that $PM_{2.5}$ lagged values keep their importance.

As concluded in [18], the climatic parameters did not improve the performance of the models when predicting over a large time horizon. Models using $PM_{2.5}$ only performed better than those using climatic parameters. On the other hand, when we used all of the climatic parameters, the model performed better than those with selected climatic parameters. Also, the lagged values of the selected climatic parameters did not show any improvement; on the contrary, they tended to worsen the prediction.

To summarize, based on our analysis, we find that for a smaller lagged value window size, Random Forest (RF) and Adaboost demonstrate the best performance. However, as the lagged values window size exceeds 22 days and above, LightGBM emerges as the most effective model. Intriguingly, the concentration of $PM_{2.5}$ from the previous day appears to be the most crucial feature. This finding can be attributed to the tendency of $PM_{2.5}$ to stagnate between two days under certain climatic conditions. The second most important feature is the concentration of $PM_{2.5}$ from the 5th previous day. This observation can be explained by the persistence of the peak from the last week, which continues to impact the $PM_{2.5}$ concentration during the peak of the following week. As illustrated in Fig. 3, the weekly seasonality of $PM_{2.5}$ results from road traffic patterns, leading to occasional peaks occurring on Monday and Tuesday, while Fridays and Saturdays coincide with the Algerian weekend. This temporal pattern contributes to the significance of the 5th previous day's $PM_{2.5}$ concentration as a predictive feature. These findings shed light on the significant role of lagged values and specific features in accurately forecasting the weekly peak of $PM_{2.5}$. Such insights can enhance our understanding of air quality dynamics and assist in developing more effective environmental forecasting strategies.

Table 3 shows the performances of the proposed models and models from related works, specifically those designed to predict PM_{10} and $PM_{2.5}$ in Algiers and cities with similar climatic conditions. It is worth mentioning that this comparison aims to show how the proposed models perform and not to compare between the models, since each one is designed using different data concerning different periods and cities. In terms of R^2, the [7] model outperforms our model. However, it has only been designed and tested using 4 months of data, and did not include the seasonality aspect of $PM_{2.5}$. In terms of RMSE model described in [19] performed similarly to our model. Other study shows a better performance than our model [10].

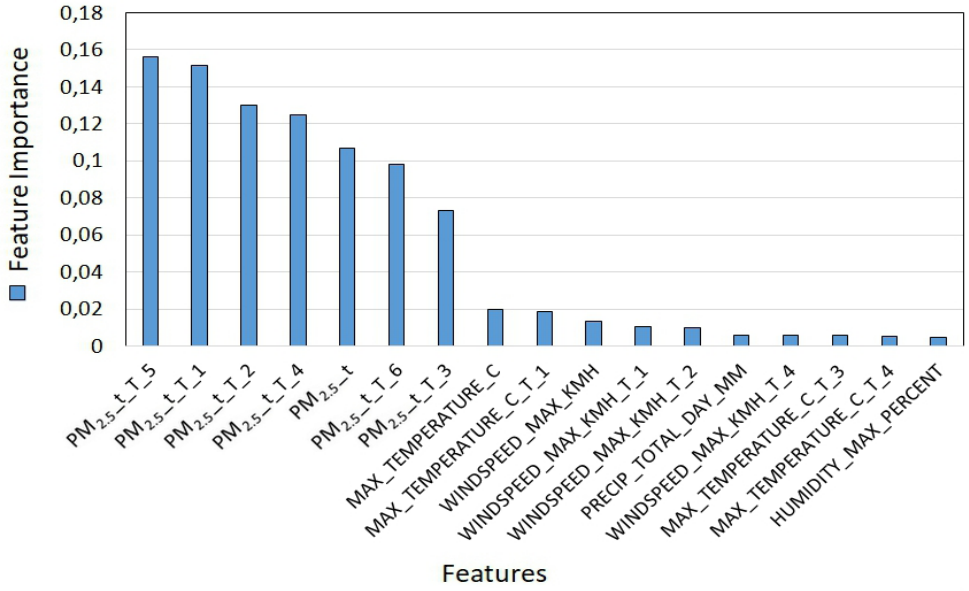

Fig. 9. The RF features importance model using selected climatic parameters and $PM_{2.5}$, 7 day lagged values

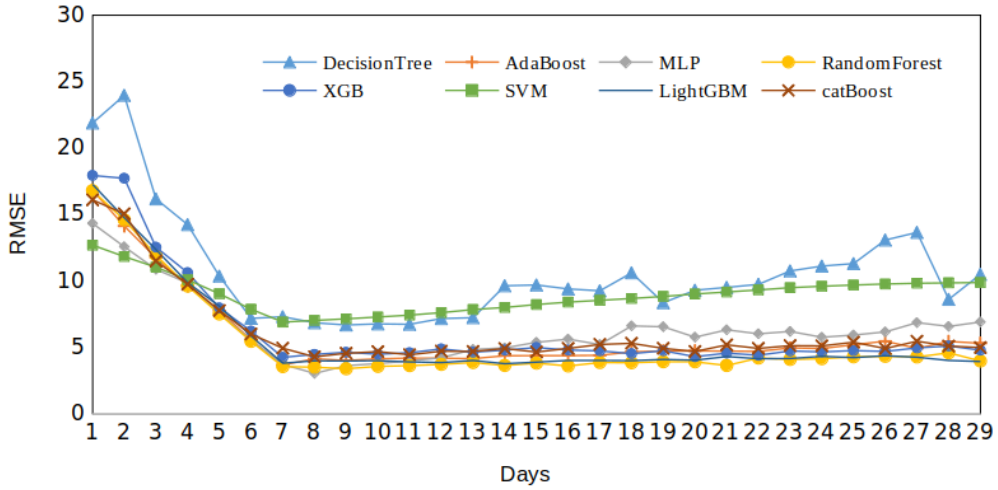

Fig. 10. RMSE evolution according to the size of the lagged values window

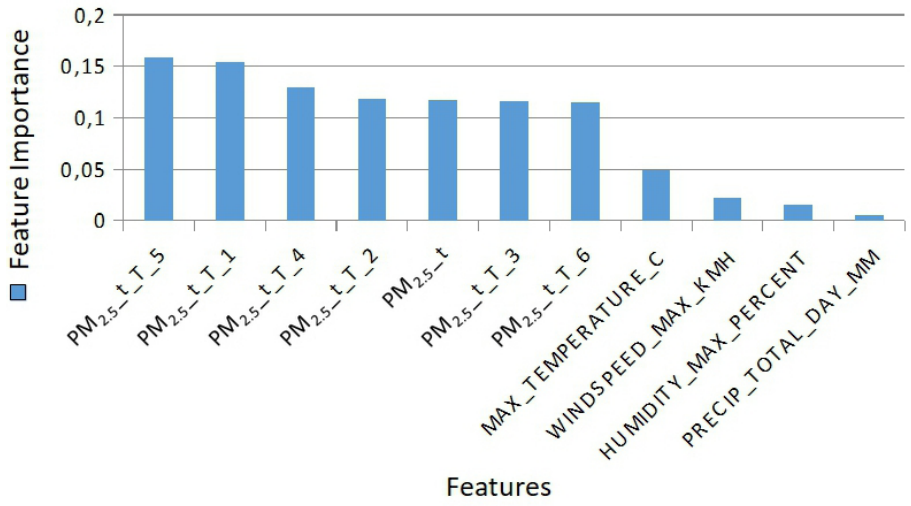

Fig. 11. RF features importance using 7 day lagged values of $PM_{2.5}$ and no lagged values of climatic parameters

Conclusion

Efficient prediction of $PM_{2.5}$ peaks can be achieved with models that do not require expensive computing power. The model proposed in this paper is easily designed, deployable and can be integrated into the decision-making process. Unlike deep learning models, machine learning models are often considered more interpretable and offer the possibility to inspect the importance of each feature for the model output. The study focused on Algiers, North Algeria, where road traffic was found to be the primary source of pollution. Weekly seasonality was confirmed, and this was utilized to improve the prediction accuracy of the proposed model. The quality and reliability of the proposed models were evaluated using statistical metrics such as RMSE, MAE, and R^2. Ensemble learning models were found to accurately forecast $PM_{2.5}$ peaks, with feature selection methods significantly impacting model outcomes. The use of lagged values with a window size of multiples of seven significantly reduced the model's prediction error. The Adaboost model performed the best when using only $PM_{2.5}$ and a 7-day lagged value. RF outperformed other models except for input combinations with large input size, where lightGBM outperformed RF. The use of lagged values of climatic parameters did not improve the performance, as changes in climatic parameters do not immediately affect the weekly peak of $PM_{2.5}$ concentration. The model with selected climatic parameters and only $PM_{2.5}$ lagged values showed better performance than those using lagged values of both climatic parameters and $PM_{2.5}$ values. The built-in feature importance of the random forest model confirmed that the lagged values of $PM_{2.5}$ are more important than climatic parameters, even those selected according to their correlation with the $PM_{2.5}$. Future work includes incorporating data on road traffic, emission source, and optical aerosol depth, as well as visualizing pollution dispersion in the geographic area to aid decision-makers in managing peak periods.

Financial supports

The study reported in this paper did not receive any funding.

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.

Acknowledgements

The authors would like to thank all those who contributed to this research.

Ethical considerations

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely ob-served by the authors.

References

1. Bouhila Z, Mouzai M, Azli T, Nedjar A, Mazouzi C, Zergoug Z, Boukhadra D, Chegrouche S, Lounici H. Investigation of aerosol trace element concentrations nearby Algiers for environmental monitoring using instrumental neutron activation analysis. Atmospheric Research. 2015 Dec 1;166:49-59.

2. Talbi A, Kerchich Y, Kerbachi R, Boughedaoui M. Assessment of annual air pollution levels with PM_1, $PM_{2.5}$, PM_{10} and associated heavy metals in Algiers, Algeria. Environmental Pollution. 2018 Jan 1;232:252-63.

3. Belarbi N, Belamri M, Dahmani B, Benamar MA. Road traffic and PM_{10}, $PM_{2.5}$ emission at an urban area in Algeria: identification and statistical analysis. Pollution. 2020 Jul 1;6(3):651-60.

4. Ladji R, Yassaa N, Balducci C, Cecinato A. Particle size distribution of n-alkanes and polycyclic aromatic hydrocarbons (PAHS) in urban and industrial aerosol of Algiers, Algeria.

Environmental Science and Pollution Research. 2014 Feb;21:1819-32.

5. Pu Q, Yoo EH. Ground $PM_{2.5}$ prediction using imputed MAIAC AOD with uncertainty quantification. Environmental Pollution. 2021 Apr;274:116574.

6. Chellali MR, Abderrahim H, Hamou A, Nebatti A, Janovec J. Artificial neural network models for prediction of daily fine particulate matter concentrations in Algiers. Environmental Science and Pollution Research. 2016 Jul;23:14008-17.

7. Ibrir A, Kerchich Y, Hadidi N, Merabet H, Hentabli M. Prediction of the concentrations of PM_1, $PM_{2.5}$, PM4, and PM_{10} by using the hybrid dragonfly-SVM algorithm. Air Quality, Atmosphere & Health. 2020 Sep 11;14(3):313–23.

8. Wang Y, Wang H, Zhang SH. Prediction of daily $PM_{2.5}$ concentration in China using data-driven ordinary differential equations. 2020 Jun 15;375:125088–8.

9. Wu H, Liu H, Duan Z. $PM_{2.5}$ concentrations forecasting using a new multi-objective feature selection and ensemble framework. Atmospheric Pollution Research. 2020 Jul;11(7):1187–98.

10. Liou NC, Luo CH, Mahajan S, Chen LJ. Why is Short-Time $PM_{2.5}$ Forecast Difficult? The Effects of Sudden Events. IEEE Access. 2020;8:12662–74.

11. Analitis A, Barratt B, Green D, Beddows A, Samoli E, Schwartz J, et al. Prediction of $PM_{2.5}$ concentrations at the locations of monitoring sites measuring PM_{10} and NOx, using generalized additive models and machine learning methods: A case study in London. Atmospheric Environment. 2020 Nov;240:117757,

12. Xing H, Wang G, Liu C, Suo M. $PM_{2.5}$ concentration modeling and prediction by using temperature-based deep belief network. Neural Networks. 2021 Jan 1;133:157-65.

13. Harishkumar KS, Yogesh KM, Gad I. Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models. Procedia Computer Science. 2020 Jan 1;171:2057-66.

14. Kamińska JA. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wrocław. Journal of environmental management. 2018 Jul 1;217:164-74.

15. Miskell G, Pattinson W, Weissert L, Williams D. Forecasting short-term peak concentrations from a network of air quality instruments measuring $PM_{2.5}$ using boosted gradient machine models. Journal of environmental management. 2019 Jul 15;242:56-64.

16. Gao X, Li W. A graph-based LSTM model for $PM_{2.5}$ forecasting. Atmospheric Pollution Research. 2021 Sep 1;12(9):101150.

17. Ma J, Ding Y, Cheng JC, Jiang F, Gan VJ, Xu Z. A Lag-FLSTM deep learning network based on Bayesian Optimization for multi-sequential-variant $PM_{2.5}$ prediction. Sustainable Cities and Society. 2020 Sep 1;60:102237.

18. Zhang B, Zhang H, Zhao G, Lian J. Constructing a $PM_{2.5}$ concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. Environmental Modelling & Software. 2020 Feb 1;124:104600.

19. Pak U, Ma J, Ryu U, Ryom K, Juhyok U, Pak K, Pak C. Deep learning-based $PM_{2.5}$ prediction considering the spatiotemporal correlations: A case study of Beijing, China. Science of the Total Environment. 2020 Jan 10;699:133561.

20. Xu X, Tong T, Zhang W, Meng L. Fine-grained prediction of $PM_{2.5}$ concentration based on multisource data and deep learning. Atmospheric Pollution Research. 2020 Oct 1;11(10):1728-37.

21. Hough I, Sarafian R, Shtein A, Zhou B, Lepeule J, Kloog I. Gaussian Markov random fields improve ensemble predictions of daily 1 km $PM_{2.5}$ and PM_{10} across France. Atmospheric Environment. 2021 Nov 1;264:118693.

22. Stafoggia M, Bellander T, Bucci S, Davoli M, de Hoogh K, de' Donato F, et al. Estimation

of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environment International [Internet]. 2019 Mar 1;124:170–9. Available from: https://www.sciencedirect.com/science/article/pii/S0160412018327685.

23. Zamani Joharestani M, Cao C, Ni X, Bashir B, Talebiesfandarani S. $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere. 2019 Jul 4;10(7):373.

24. Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.

25. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences. 1997 Aug 1;55(1):119-39.

26. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016 Aug 13 (pp. 785-794).

27. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30.

28. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems. 2018;31.
29. Algerian Office of Statistics [Internet]. [Cited 19-07-2021]. 2018. Available from: .https://www.ons.dz/IMG/pdf/Demographie. 2018.pdf.

30. Algerian Office of Statistics [Internet]. [Cited 31-10-2021]. 2019. Available from: https://www.ons.dz/IMG/pdf/e.immats2_2019.pdf

31.Project TWAQI. Air Quality Historical Data Platform [Internet]. Available from: https://aqicn.org/data-platform/

32. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Computational Statistics & Data Analysis. 2015 Oct 1;90:84-99.

33. Zhou Y, Chang FJ, Chang LC, Kao IF, Wang YS, Kang CC. Multi-output support vector machine for regional multi-step-ahead $PM_{2.5}$ forecasting. Science of the Total Environment. 2019 Feb 15;651:230-40.