Ensemble Learning Models for the Prediction of the Weekly Peak of PM2.5 Concentration in Algiers, Algeria

Sabri Ghazi 1, Ahmed Dib 2, Mehdi Mendjel 1, Khadir Tarek 1, Julie Dugdale 3

1,3 Laboratoire de Gestion Electronique de Documents, Department of Computer Science, University Badji Mokhtar, P.O. Box 12, 23000, Annaba, Algeria

2 Laboratoire Réseaux et Systèmes, Department of Computer Science, University Badji Mokhtar, P.O. Box 12, 23000, Annaba, Algeria

3 University Grenoble Alps, LIG, France

E-mails: 1sabri.ghazi@univ-annaba.dz

**Highlights**:

* The use of lagged values of PM2.5 improves the performances of the models in predicting the weekly peak of PM2.5.
* The lagged values of PM2.5 are the most important according to the built-in features importance of the Random Forest model
* Due to the weekly seasonality of the road traffic in Algiers, lagged values with a window size of seven and multiples of seven days show the best performances.
* When using daily measures of PM2.5 and climatic parameters, lagged values of climatic parameters worsen the prediction performances of the weekly PM2.5 peak, even those selected according to their correlation with PM2.5.

**Abstract**:

ML (Machine Learning) models are commonly used to predict PM2.5 (Particulate Matter with an aerodynamic diameter less than 2.5 micrometre) concentration. However, their performances depend on the used FS (Features Selection) method and the values of the hyperparameters. Despite their impact on models performances, FS and lagged values are frequently ignored in studies predicting air pollution concentration. This paper aims to assess the performance of ML models to predict the weekly peak of PM2.5 concentration in the city of Algiers, Algeria. This, using different features combinations and lagged values windows’ sizes. We used SVM (Support Vector Machine), MLP (Multi-Layers Perceptron), DT (Decision Tree), and six ensembles models, specifically: AdaBoost ( Adaptive Boosting), XGboost (eXtreme Gradient Boosting), LightGBM (Lights Gradient Boosting), CatBoost, and RF (Random Forest). The used dataset includes three years of daily measures of twenty-one weather parameters and PM2.5 concentration. We found that lagged values of PM2.5 improve the performances, specifically, when the size of lagged values window is seven days or a multiple of seven days. Consequently, we confirmed that road traffic, which has a weekly seasonality, is the primary source of PM2.5 in Algiers. The study points out that lagged values of weather parameters worsen the prediction performances, even those selected using their correlation with PM2.5. The Adaboost model presented the best performance, its RMSE (Root Mean Squared Error) is 2.899 µgram/m3 and its R2 is 0.96, followed by MLP and RF. The presented model outperformed the state-of-the-art models, specifically those designed to predict the concentration of PM2.5 in Algiers.

**Keywords** PM2.5, Air pollution, Ensemble Learning, Time series forecasting, Algiers, Air pollution prediction.

**Introduction**

The degradation in air quality has emerged as a major challenge facing many cities in the world. In the developing countries, uncontrolled urban expansion, fossil energy-based transportation, and the lack of legislation to enforce air quality standards, lead to a very alarming air pollution levels. Peak period occurs when the concentrations of air pollutants are above the tolerated level. Among the pollutant responsible for peaks is the PM2.5, it 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 source of PM2.5 spread. (Bouhila et al., 2015) studied the heavy metal content of PM2.5 in Algiers; it determined that Fe and Sc are highly present and concluded that the annual level of PM2.5 is beyond local and international standards. The same conclusion is confirmed in (Talbi et al., 2018), in which authors analysed samples of PM1, PM2.5, and PM10 from two stations in Algiers during 2015 and 2016 in an urban and roadside. By inspecting the samples of PM2.5, the concentrations of heavy metals were determined, with Pb representing 5%. (Belarbi et al., 2020) studied the composition of PM10 and PM2.5 in an urban area in Algiers. The heavy metal content of PM2.5 confirmed their origin from road traffic and Saharan dust. It is worth to mention that Algeria is the only country on the planet to continued using leaded carburant, untill August 2021 when the Algerian government enacts a law banning the use and sale of leaded carburant. Due to its diameter and toxicity, PM2.5 can be inhaled by human being and leads to a serious health problem (Ladji et al., 2014). Therefore, having an accurate prediction of PM2.5 peak period can help decision-makers mitigating the crisis and reduce its effects, specifically by alarming the population with special medical conditions. Many approaches have been used to predict PM2.5 concentration, they can be categorized into five categories: Deterministic Models, Linear models, Machine Learning based models, hybrid models, Satellite-derived Aerosol Optical Depth model (Pu & Yoo, 2021). Moreover, one may categorise them according to: the model inputs; the prediction horizon and the studied region. A non-exhaustive review of recent studies proposing models to predict PM2.5 concentration is summarized in Table 01. Where, weather parameters such as Wind Speed, Wind Direction, Relative Humidity, Pressure, Ambient temperature and Cumulative precipitation are noted respectively *WS, WD, RH, Pr, T* and *P*. Anthropogenic event data is noted A. Despite its impact on the city air quality and due to the lack of measures about PM2.5 concentration in Algiers, there is limited number of studies presenting model to forecast PM2.5 in Algiers. (Chellali et al., 2016) presents a MLP model to predict the long-term concentration of PM10 at Algiers. It is trained using two years dataset of PM10 concentration and meteorological parameters (wind speed, relative humidity, and temperature), selected based on their correlation with PM10. The used dataset is relatively old, dated 2003-2004 which did not reflect the climatic and the anthological changes occurred during the recent decencies in Algiers. (Ibrir et al., 2021) proposes a SVM model to predict the concentration of PM of different sizes including PM2.5 in Algiers. To select the best model hyperparameters, authors used a swarm algorithm called Dragonfly. The described model showed relatively convenient performances. However, the used dataset is limited and covering only four months and does not include the yearly seasonality of PM2.5. Therefore, it leads to a poor generalization of the model. (Wang et al. 2020) uses an Ordinary Differential equation to model PM2.5. The model was compared with AR model, it showed relatively similar performances. However, the model was trained using a restrained dataset covering only two months of daily PM2.5 concentration, which could lead to a very weak generalization. Machine learning model are commonly used and compared with linear models. In (Wu et al., 2020) the PM2.5 times series is smoothed using Wavelet transformation in order to eliminate short-term fluctuation, which impacts the accuracy of the prediction. To avoid the effects of the sudden change, (Liou et al., 2020) used an unsupervised method to cluster anthropogenic and environmental events. As described, the unexpected event such as rainfall intensity, wind speed, and road traffic have an impact on the concentration of PM2.5. Events data are collected from the error in the forecast of an Adaptive Iterative Forecast model. To tackle the lack of PM2.5 measurement in London, (Analitis et al. 2020) developed a PM2.5 concentration prediction model. The model uses the concentration of PM10 and NO as inputs. Linear regression and Random Forest models are combined using GAM (Generalized Additive Model). The authors tested many combinations with weather parameters to get the best-performing model. To predict PM2.5 in Beijing, China, (Xing et al. 2021) used TDBN (Temperature-Nased Deep Belief Networks) with many hidden layers and different size. (Doreswamy et al., 2020) included topographical data among the inputs, and compared ML models to forecast PM2.5 in Newport, Taiwan. The performance of RF in the prediction of PM2.5 was investigated in (Kamińska, 2018) . The dataset is divided into many subsets, and assessed the accuracies in each one, concluding that RF is more accurate to predict PM2.5 in warmer periods. (Miskell et al., 2019) adopted a binary classification approach to predict PM2.5 exceeds. The PM2.5 measures are converted into two class : Peak and No-Peak. Though, the number of peaks is always less than the normal level, this results as an imbalanced dataset, since peaks is a minority class, and thus affects the model generalization. Recent studies are proposing deep learning models of different architectures to prediction PM2.5. In (Gao & Li, 2021) many monitoring stations in Gansu, China, are modelled as a weighted graphs with LSTM nodes each one. The weight in the edge between two stations is included in the LSTM input of the other station. The model can forecast PM2.5 concentration in every station without the need to build a model for each station. According to the study, the model took into consideration the spatiotemporal information, and therefore performed better than the ensemble learning model, using the same dataset. (Ma et al., 2020) a Bayesian optimization is used to determine the values of the hyperparameters 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 showed the best performances. However, the used data to validate the model was randomly selected, in time series data this could lead to poorly explicative data since it lacks the time order of each observation. In (Zhang et al., 2020) , authors used an Auto-Encoder to compress the feature space before passing it as input to a LSTM. The proposed model receives as input the lagged values of PM2.5, snow, precipitation, ambient temperature, wind speed, and direction. Compared to classic models such as CAMx, CMAQ, and other deep learning models, the proposed model showed the best performances. Authors argue that for a long-term prediction the model trained using only PM2.5 performed better than the one that includes weather parameters. Nevertheless, for a small prediction horizon, the models with those parameters showed better precision. (Pak et al. 2020) Mutual Information Estimator is used for determining the correlation between times series of weather pollutant parameters from 384 stations across China. Authors claim that this helps to capture spatiotemporal information. The selected features are then used to train CNN-LSTM model. (Xu et al. 2020) used multi-stages method in order to consider spatial and temporal information in the prediction of PM2.5. Initially, in each monitoring station using LSTM, a spatial predictor and a temporal predictor are trained using the appropriate data. Secondly, the output of each is used in a Regression Tree model to predict PM2.5 concentration, and lastly, an ANN is used to predict a grid level PM2.5. Some studies included additional inputs such as AOD, (Hough et al. 2021) used AOD and empirical data to predict the daily PM10 and PM2.5 concentrations in France. A RF model is used to impute PM2.5 in PM10 only stations. Missing values from AOD are also predicted using an RF model. GAM is used to combine the output of Linear Regression, RF and GMRF (Gaussian Markov Random Field). The same strategy is used in (Stafoggia et al., 2019) to predict PM10 and PM2.5 in Italy, with adding at the last stage a local predictor to improve the prediction at a small scale. Using data about Tehran city, Iran, (Zamani Joharestani et al., 2019) investigated the contribution of AOD in enhancing the performances of PM2.5 prediction model. Ensemble Learning models are easy to implement and require less computing, and explainable compared to Deep Learning models. The aim of this paper is to assess the performance of ensemble learning models to predict the weekly peak of PM2.5. Furthermore, this paper aims to determine the impact of FS and the used lagged values, commonly ignored in studies presenting PM2.5 prediction models. In order to compare the EL models performance, classic machine learning models are also used such as MLP, SVM.

The remaining of this paper is organized as follows: The next section presents a description of the ensemble learning and gradient descent approachs. The Material and the Method section presents the studied region, it describes the used dataset and its statistics properties. The results are described and discussed in Section 4, in which a comparison between different types of models is presented; furthermore it presents a comparison between the presented models performances and related works models’ performances. Finally, in [Conclusions and Future Work](https://journal-bcs.springeropen.com/articles/10.1186/s13173-017-0052-0#Sec9) section we conclude and suggest some possible perspectives to this work.

**Background on Ensemble Learning approach**

A Decision Tree is a machine-learning model which 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 split the samples. DT is commonly used in Ensemble Learning approach, 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 affects an equal weight to each model output in the vote to determine the final output. In a Random Forest (Breiman, 2001); it uses decision tree models and combines their outputs. Many decision tree models are trained using random samples of the training data and random subsets of the features. AdaBoost algorithm, short for adaptive boosting (Freund & Schapire, 1997) is a ensemble learning boosting technique. It determes the parameters by re-assigning the weights to each instance, with higher weights to incorrectly classified instances. XGBoost (Chen & Guestring, 2016) is a decision-tree-based ensemble learning that uses a gradient boosting steroid strategy. It applies the principle of boosting and provides a parallel tree boosting. LightGBM is a recent improvement of the gradient boosting algorithm (Ke et al., 2017). 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 rather than leaf-wise. Therefore, when growing on the same leaf in LightGBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy, which can rarely be achieved by any of the existing boosting algorithms. Another version of gradient descent is CatBoostGBM (Prokhorenkova, et al. 2017), it is a gradient descent algorithm designed to deal with categorical featuers and also avoid the overfitting problem.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Area and Period** | **Prediction horizon** | **Model** | **Features Engineering & hyperparameters** | **Lagged values** | **Inputs** | **Multi /single output** |
| (Chellali et al., 2016) | Algiers, 2003-2004 | 24 h | MLP | Correlation | - | PM2.5,WS, RH, T | single |
| (Ibrir et al., 2021) | Algiers, 4 months | Not mentioned | SVM | Correlation, Dragonfly | - | PM2.5,WS, RH, T, P | Single |
| (Liou et al., 2020) | Taichung ,Taiwan, 2017 | 3 h | AIF | Hierarchical Clustering | - | WS, RH,T, P, A | Single |
| (Gao & Li, 2021) | Gansu, China, 2019-2020 | From 1h to 48h | G-LSTM | Adjacency Matrix | 4 h | PM2.5, WS, WD, RH, T, P, Pr, CO,NO2, O3,SO2 PM10, PM2.5 | Both |
| (Ma et al., 2020) | Wayne ,Michigan, USA | - | Lag-FLSTM | Bayesian optimization | 48 h | PM2.5, WS, WD, Press, T, CO, SO2, NO2, PM10 | Single |
| (Zhang et al., 2020) | Beijing, China | From 24h to 168h | AE-Bi-LSTM | Auto-Encoder | - | PM2.5, WS, P, Snow, T, Dewpoint | Single |
| (Pak et al., 2020) | Beijing, China, 3 years 2015-2017 | - | CNN-LSTM | Mutual Information estimator | - | CO, SO2, O3, NO2, PM2.5, PM10, T, WD, WS, | Single |
| (Xu et al., 2020) | Beijing-Tianjin-Hebe, China | From 1h to 24h | LSTM | Regression Tree, ANN | - | WS, WD, RH, T, Workday/Weekend, Pres,P,, Dew point, Season, Month, PM2.5 | Single |
| (Wang et al., 2020) | 2015-2016 | - | Ordinary Differential Equation | Genetic Algorithm | - | PM2.5 | Single |
| (Analitis et al., 2020) | London, UK, 2004-2013 | 1h | Linear Regression, Random Forest | Generalized Additive Model | - | NO2, PM10, PM2.5, Latitude, T, Week day, WS, WD, RH, Roadside vs Background | Single |
| (Hough et al., 2021) | France, 2000-2019 | 1 day | Gaussian Markov Random Field,Random Forest, | Generalized Additive Model | - | PM10, PM2.5, AOD, P, T, WS. | Single |
| (Stafoggia et al., 2019) | Italy, 2013-2015 | 1 day | Random Forest | - | - | AOD, PM2.5, PM10, WD, WS, Press, P, T | Single |
| (Zamani Joharestani et al., 2019) | Iran, Tehran, 2015-2018 | 1 day | Random Forest, XGBoost | RF Features Importance  XgBoost Features Importance  Permutation Importance. | 2 Days | AOD, PM2.5, WS, RH,WD, P, Press, T, Dew Point | Single |
| (Xing et al., 2021) | Beijing, China, 2018 | 1 day | Temperature-Nased Deep Belief Networks | - | - | WS, P, T, PM10, SO2, CO2, Pess, RH | Single |
| (Doreswamy et al., 2020) | Newport ,Taiwan, 2012-2017 | - | XGBoost, RF, MLP, Decsion Tree, K neares neighbours | - | - | - | Single |
| (Kamińska, 2018) | Wrocław, Poland, 2015-2016 | 1 hour | RF | RF Features Importance. | - | Road Traffic, T, WS, WD, RH, Press, week day, holidays, month. | Single |
| (Miskell et al. 2019) | Christchurch, New Zealand | 1 hour peak  1 day peak | boosted gradient machine | - | - | T, WS,NO, NO2 | Single |
| (Wu et al., 2020) | Hohhot, Harbin, Wuhan, Changsha China | 1 hour , 2 hour, and 3 hour | outlier robust extreme learning machine | -nonconvex sparse regularization  -wavelet transform | - | - | Multi |
| (Chang-Hoi et al., 2021) | Korea, Seoul | 1 day , 2 days | - | - | - | - | single |

**Table 01: Studies presenting PM2.5 concentration prediction**

**Material and methods**

**Studied Region**

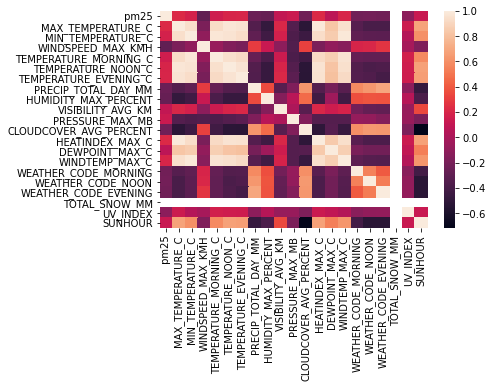
Algiers is located in the Centre of the North of Algeria, it is a coastal city bordered by the Mediterranean Sea on the North. It is the economic and the political capital of Algeria. According to the Office National of Statistics (ONS, 2018), in 2019 the estimated population was 8 million habitants. The city has a high economic attraction; it hosts many central administrations, international corporations’ headquarters, and four active industrial zones. The public transportation in Algiers did not expand proportionally to the demography; therefore, the transportation is mainly based on personal cars, according to (ONS, 2019) Algiers’s motor fleet reached 2 million in 2019. The city contains a seaport in which goods are mainly transported using trucks.

**Dataset Description**

This study uses a dataset covering 3 years (from 2019 to 2021) of daily measures of climatic parameters and PM2.5 concentration. The measures of PM2.5 are collected by EPA US-EMBACY station in Algiers, GPS coordination are 36.75595300548415; 3.039189599146588. The climatic parameters are provided by official meteorology agency (ONM). Table 02 describes some statistics properties. Some important events occurred during the period of the dataset, the first is the COVID-19 lockdown, which started from March 2020 to December 2020, and also during the second peak during August 2021. Moreover, the forest fires in the Tizi-Ouzou mountains which lasted for 7 days, from 9 to 15 August 2021. Figure 1 illustrates a positive and negative correlation between PM2.5 and the climatic parameters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Mean** | **std** | **min** | **max** | **Missing value** |
| PM2.5 | 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% |
| WINDSPEED\_MAX\_KMH | 16.46 | 6.75 | 4,00 | 44.00 | 0% |
| TEMPERATURE\_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% |
| CLOUDCOVER\_AVG\_(%) | 28.80 | 25.28 | 0.00 | 94.375 | 0% |
| HEATINDEX\_MAX (C°) | 24.13 | 6.66 | 10.00 | 44.00 | 0% |
| DEWPOINT\_MAX (C°) | 14.46 | 4.79 | 2.00 | 26.00 | 0% |
| WINDTEMP\_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% |
| SUNHOUR | 10.48 | 3.02 | 3.5 | 14.5 | 0% |

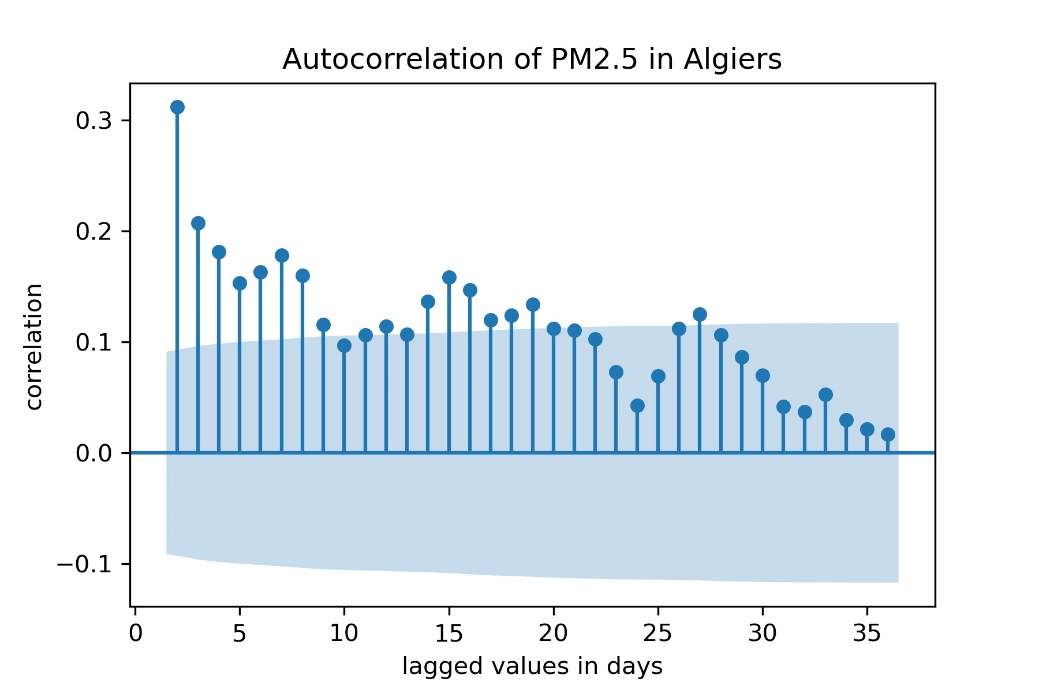
**Table 2: statistical properties of the dataset**



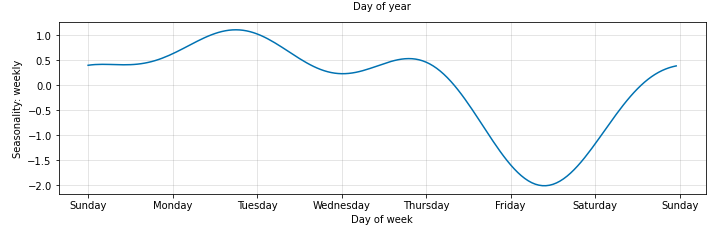
**Figure 01: Correlation between the features of the dataset**

**Auto-correlation**:

Figure 2 shows the auto-correlation of PM2.5, which measures the correlation between the lagged values and the current value of the PM2.5 time series. The lagged values are varied from 1 day to 35 days. As illustrated, the local peaks mean the positive correlation between the value of the PM2.5 and its past values, specifically those of days numbers are multiples of seven such as 7, 14, 21, 28. This is also confirmed in Figure 3, in which the weekly seasonality is clearly shown. This shows how PM2.5 concentration decreases during Friday and Saturday, which are the Algerian local weekend holiday. During the weekday the PM2.5 concentration increases specifically, on Sunday, Monday, Tuesday and Thursday. With a local peak on Tuesday.



**Figure 2: Auto-correlation of PM2.5 time series.**



**Figure 3 : Weekly seasonality of PM2.5 in Algiers**

**Data normalisation and missed values imputation:**

As described in Table 02, PM2.5 times series contains 10% of missed values. To maintain the time order and its impact, we imputed them using KNN (K Nearest Neighbours) imputer. This algorithm (Tutz, & Ramzan, 2015) uses a Euclidian distance to determine the *K* closest complete samples of the dataset. Then it fills the missed values with a weighted average of the neighbours. Since the features are in different scales, we normalized the data using equation (1).

(1)

Where *min* and *max* are function which compute the minimum and maximum value.

**Performances metrics:**

In order to compare the performances of the models, we used RMSE (Root of mean Square error), MAE (Mean Absolute Error) , and R2, as defined in **(2)** , **(3)**, **(4),** correspondingly.

(2)

(3)

(4)

Where *yi, measured* is the *ith* measured value of a vector of *n* values, yi, predicted is the *ith* predicted value of the vector of *n* values. is the mean of the measured value.

**Results and discussion**

The objective is to design a model which maps the input *PM2.5(t),PM2.5(t-1)…PM2.5(t-k), WeatherFactor1(t),WeatherFactor1(t-1)….WeatherFactor1(t-k),...WeatherFactorm(t),WeatherFactorm(t-1),… WeatherFactorm(t-k)* to the output representing the peak of the next week: *max(PM2.5(t+1),PM2.5(t+2),PM2.5(t+3),PM2.5(t+4),PM2.5(t+5),PM2.5(t+6),PM2.5(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, *max* is a function that returns the maximum values of PM2.5. To train the models we used the first 70% of the dataset, where the remaining 30% are used to test the performance of the models. We computed the peak of each week of the dataset to form the target varaible.

**Models using all features**

Models receive as inputs all the features of the dataset, 21 climatic features, and PM2.5. Different models are trained using lagged values from one day to 30 days. As described in figure 03, the best performing model is RF with RMSE of 3.648 and R2 of 0.937, for lagged values of 7 days. The next best-performing model is AdaBoost with RMSE of 4.770 and R2 of 0.892. The order changes with lagged values of 24 days, 28 days, 29 days, and 30 days, in which LightGBM overperformed RF. For example, LightGBM shows RMSE of 4.566 and R2 of 0.901 where RF is 4.832 and R2 is 0.889 for lagged values of 24 days. Figure 3 shows the evolution of the RMSE according to the number of lagged values. Except for SVM, the others models' performances start to improve when inputs with seven days lagged values are used. Figure 4 shows the features importance determined using RF built-in method. The features X ist j lagged values is noted X\_t\_T\_j. PM2.5 lagged values come first specifically PM2.5(t-1), PM2.5(t-5), PM2.5(t-2), PM2.5(t-3), PM2.5(t-4), after that it comes the first climatic parameters Pressures\_Max. As illustrated in Figure 4, the lagged values of climatic parameters are not considered important in the RF model with 7 days lagged values, which is the best performing model.

**Figure 03: The RMSE evolution according to the number of lagged values, models using all the features.**

**Figure 4: The RF features Importance using 7 days lagged values, all features included.**

**Models using PM2.5 only:**

Those models receive as input only PM2.5 lagged values, no climatic parameters are used. We varied the lagged values windows’ size from 1 day to 30 days. Models with inputs with lagged values of 7 days, the best performance is shown by Adaboost with an RMSE of 2.899 and R2 of 0.960 , followed by MLP RMSE 2.915 and an R2 of 0.959, RF shows an RMSE of 2.918 and R2 of 0.959. This means that when using the lagged values of the time series only, MLP and Adaboost perform best. The order changes for lagged values of 21, 22, 24,25, 26, 27, 28, 29, and 30, in which LightGBM shows the best performances, for example, LightGBM with 25 days lagged values gives an RMSE of 3.791. The best Model with 23 lagged values is RF with an RMSE of 3.888. Figure 5, shows the RMSE of the models trained using input with lagged values from 1 day to 30 days. Figure 06 shows the features importance of the RF model trained using 7 lagged values, lagged values :PM2.5(t-1), PM2.5(t-5) seams to keep their importance.

**Figure 05: The RMSE according to the number of lagged values of PM2.5, models using only PM2.5 no climatic parameters**

**Figure 6: Features Importance of a RF model trained only with PM2.5, with 7 Days lagged values**

**Model using PM2.5 and features selected with correlation:**

Those models are trained using inputs of PM2.5 and selected climatic parameters. The selection of the parameters is done using correlation. Among all tested combinations, models using 7 days lagged values present the best performances. The RF model shows the best performance with an RMSE of 3.791 and R2 of 0.931, followed by LightGBM RMSE of 4.345. As presented in figure 07, the order changes with 8 lagged values, the 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. Figure 08 shows the importance of the features of the best performing model, as illustrated PM2.5 lagged values come first, specifically PM2.5(t-5) and PM2.5(t-1), After that the temperature seems to be the most considered weather factor.

**Figure 7: RMSE evolution according to the size of lagged values window size.**

**Figure 8, The RF features importance model using selected climatic parameters and PM2.5, 7 days lagged values.**

**Models with selected features and only lagged values of PM2.5**

We trained the models using an input composed of lagged values of PM2.5 and no lagged values of the selected weather parameters. This is done in order to determine how much the lagged values of weather factors can impact the models' performances. As shown in figure 9, the best performing model is found to be the MLP for 8 days lagged values, it shows an RMSE of 3.039, followed by RF with an RMSE of 3.505.

The order changes with 9 lagged values, the latter shows an 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 an RMSE of 3.934. Figure 10 illustrates the importance of the features, and PM2.5 lagged values keep their importance.

**Figure 9: RMSE evolution according to the size of lagged values window size.**

**Figure 10: RF features importance using 7 days lagged values of PM2.5 and no lagged values of climatic parameters.**

As concluded in (Zhang et al., 2020), when predicting for a large time horizon, the climatic parameters did not improve the performance of the models. Models using PM2.5 only performed better than those using climatic parameters. On the other hand; when we used all the climatic parameters, the model performed better than those with selected climatic parameters. Also, the lagged values of the selected climatic parameters did not present any improvement, on the contrary, they tend to worsen the prediction.

Table 3 presents the performances of the proposed models and some related works model, specifically those designed to predict PM10 and PM2.5 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 period and city. In terms of R2, (Ibrir et al., 2020) model outperforms our model, however it has been only designed and tested using 4 months of data, and did not includes the seasonality aspect of PM2.5. In terms of RMSE (Pak et al., 2020) performed similarly to the proposed model. (Liou & Chen, 2020) shows better performance than our model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Study** | **R2** | **RMSE** | **MAE** |
| This work- Selected Feature and ony lagged PM2.5 | 0.956 | 3.039 | 2.495 |
| This work- Lagged values of both Selected Features and PM2.5 | 0.931 | 3.791 | 2.534 |
| This work- Only PM2.5 | **0.960** | **2.899** | **1.843** |
| This work All features included | 0.937 | 3.648 | 2.551 |
| (Zamani et al. 2019) | 0.8 | 9.93 | 13.58 |
| (Zhang et al. 2020) 168 hours ahead | - | 7.93 | - |
| (Zhou et al. 2021) | - | 3.58 | 7.44 |
| (Ma et al. 2021) | - | 3.482 | 1.85 |
| (Gao & Li, 2021) | - | 3.405 | 2.60 |
| (Xu et al. 2020) | 0.87 | 24.24 | 8.25 |
| (Pak et al. 2020) | - | 2.870 | 2.11 |
| (Chellali et al. 2016) | 0.85 | 13.780 | - |
| (Xing et al 2021) | 0.86 | 11.190 | 12.29 |
| (Ibrir et al. 2020) | 0.98 | 1.926 | - |
| (liou & chen 2020) | - | 1.780 | 1.300 |
| (Doreswamy et al., 2020) | 0.89 | - | - |
| (Stafoggia et al. 2019) | 0.81 | 5.36 | - |
| (Analitis et al. 2020) | 0.95 | - | - |
| (Kaminska, 2018) | 0.57 | - | - |
| (Chang-Hoi et al., 2021) | 0.81 | - | 5.1 |

**Table 3, The performances of the best models and some related works models.**

**Conclusion**

With the availability of the measures of PM2.5 concentration. Pearks of PM2.5 could be efficiently predicted with models with no need to expensive computing power. In contrast to deep learning model, the models described in this paper can be easily designed and deployed in developing countries and included whithin decision making processes. This paper presented a model designed using data about Algiers, North of Algeria. The road traffic appeared to be the most important source of pollution. This study confirmed its weekly seasonality, moreover we exploited this to improve the prediction accuracy of the proposed model. The quality and reliability of the proposed models is quantified using several statistical metrics :RMSE, MAE and R2. The study points out that ensemble learning models can accurately forecast PM2.5 peaks. Features selection methods have a big impact on the models outcomes. Introduced lagged values with window size of multiple of seven reduced significantly the models’ prediction error. Adaboost model presented the best performance when using only PM2.5 and its 7 days lagged values. RF dominates except for some inputs combination with large inputs size, wherein lightGBM outperformed it. The use of PM2.5 lagged values presented the best performance, and outperformed models which use lagged values of climatic parameters, this could be interpreted as the changes in climatic parameters could lead to poor models generalisation, since the effect of those changes do not immediately affects the weekly peak of PM2.5 concentration. Introducing the lagged values of climatic parameters did not improve the performances, and model with selected climatic parameters and only PM2.5 lagged values, showed better performances over those using lagged values of both climatics and PM2.5 values. In addition, the built-in feature importance of the random forest model confirmed that the lagged values of PM2.5 are more important than climatic parameters even those selected according to their correlation with the PM2.5. The use of data about road traffic, emission source and optical aerosol depth is among our perspective, furthermore visualizing the pollution dispersion among the geographic area could helps in decision making to manage the peaks period.

**Declaration of competing interest**

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

**Credit author statement author contributions**

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Sabri Ghazi, Dib Ahmed, Mehdi MENDJEL. The first draft of the manuscript was written by Sabri Ghazi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Data availability**

The datasets generated during and/or analysed during the current study are available in the github repository, <https://github.com/SabriGhazi/DataSetPM25Algiers>

**References:**

ONS, (2018) <https://www.ons.dz/IMG/pdf/Demographie2018.pdf> (Accessed 19-07-2021)

ONS, (2019) <https://www.ons.dz/IMG/pdf/e.immats2_2019.pdf> (Accessed 31-10-2021)

Analitis, A., Barratt, B., Green, D., Beddows, A., Samoli, E., Schwartz, J., & Katsouyanni, K. (2020). Prediction of PM2.5 concentrations at the locations of monitoring sites measuring PM10 and NOx, using generalized additive models and machine learning methods : A case study in London. *Atmospheric Environment*, *240*, 117757. <https://doi.org/10.1016/j.atmosenv.2020.117757>

Belarbi, N., Belamri, M., Dahmani, B., & Benamar, M. A. (2020). Road Traffic and PM10, PM2.5 Emission at an Urban Area in Algeria : Identification and Statistical Analysis. *Pollution*, *6*(3), 651‑660. <https://doi.org/10.22059/poll.2020.294710.730>

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., & Di Carlo, P. (2017). Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmospheric Pollution Research*, *8*(4), 652‑659. <https://doi.org/10.1016/j.apr.2016.12.014>

Bouhila, Z., Mouzai, M., Azli, T., Nedjar, A., Mazouzi, C., Zergoug, Z., Boukhadra, D., Chegrouche, S., & Lounici, H. (2015). Investigation of aerosol trace element concentrations nearby Algiers for environmental monitoring using instrumental neutron activation analysis. *Atmospheric Research*, *166*, 49‑59. <https://doi.org/10.1016/j.atmosres.2015.06.013>

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chellali, M. R., Abderrahim, H., Hamou, A., Nebatti, A., & Janovec, J. (2016). Artificial neural network models for prediction of daily fine particulate matter concentrations in Algiers. *Environmental Science and Pollution Research*, *23*(14), 14008‑14017. <https://doi.org/10.1007/s11356-016-6565-9>

Doreswamy, K S, H., Km, Y., & Gad, I. (2020). Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. *Procedia Computer Science*, *171*, 2057‑2066. <https://doi.org/10.1016/j.procs.2020.04.221>

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119‑139. <https://doi.org/10.1006/jcss.1997.1504>

Gao, X., & Li, W. (2021). A graph-based LSTM model for PM2.5 forecasting. *Atmospheric Pollution Research*, *12*(9), 101150. <https://doi.org/10.1016/j.apr.2021.101150>

Hong, J., Mao, F., Min, Q., Pan, Z., Wang, W., Zhang, T., & Gong, W. (2020). Improved PM2.5 predictions of WRF-Chem via the integration of Himawari-8 satellite data and ground observations. *Environmental Pollution*, *263*, 114451. <https://doi.org/10.1016/j.envpol.2020.114451>

Hough, I., Sarafian, R., Shtein, A., Zhou, B., Lepeule, J., & Kloog, I. (2021). Gaussian Markov random fields improve ensemble predictions of daily 1 km PM2.5 and PM10 across France. *Atmospheric Environment*, *264*, 118693. <https://doi.org/10.1016/j.atmosenv.2021.118693>

Ibrir, A., Kerchich, Y., Hadidi, N., Merabet, H., & Hentabli, M. (2021). Prediction of the concentrations of PM1, PM2.5, PM4, and PM10 by using the hybrid dragonfly-SVM algorithm. *Air Quality, Atmosphere & Health*, *14*(3), 313‑323. <https://doi.org/10.1007/s11869-020-00936-1>

Jena, C., Ghude, S. D., Kumar, R., Debnath, S., Govardhan, G., Soni, V. K., Kulkarni, S. H., Beig, G., Nanjundiah, R. S., & Rajeevan, M. (2021). Performance of high resolution (400 m) PM2.5 forecast over Delhi. *Scientific Reports*, *11*(1), 4104. <https://doi.org/10.1038/s41598-021-83467-8>

Kamińska, J. A. (2018). 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*, *217*, 164‑174. <https://doi.org/10.1016/j.jenvman.2018.03.094>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146-3154.

Ladji, R., Yassaa, N., Balducci, C., & Cecinato, A. (2014). Particle size distribution of n-alkanes and polycyclic aromatic hydrocarbons (PAHS) in urban and industrial aerosol of Algiers, Algeria. *Environmental Science and Pollution Research*, *21*(3), 1819‑1832. <https://doi.org/10.1007/s11356-013-2074-2>

Liou, N.-C., Luo, C.-H., Mahajan, S., & Chen, L.-J. (2020). Why is Short-Time PM2.5 Forecast Difficult? The Effects of Sudden Events. *IEEE Access*, *8*, 12662‑12674. <https://doi.org/10.1109/ACCESS.2019.2963341>

Ma, J., Ding, Y., Cheng, J. C. P., Jiang, F., Gan, V. J. L., & Xu, Z. (2020). A Lag-FLSTM deep learning network based on Bayesian Optimization for multi-sequential-variant PM2.5 prediction. *Sustainable Cities and Society*, *60*, 102237. <https://doi.org/10.1016/j.scs.2020.102237>

Miskell, G., Pattinson, W., Weissert, L., & Williams, D. (2019). Forecasting short-term peak concentrations from a network of air quality instruments measuring PM2.5 using boosted gradient machine models. *Journal of Environmental Management*, *242*, 56‑64. <https://doi.org/10.1016/j.jenvman.2019.04.010>

Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations : A case study of Beijing, China. *Science of The Total Environment*, *699*, 133561. <https://doi.org/10.1016/j.scitotenv.2019.07.367>

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516*.

Pu, Q., & Yoo, E.-H. (2021). Ground PM2.5 prediction using imputed MAIAC AOD with uncertainty quantification. *Environmental Pollution (Barking, Essex: 1987)*, *274*, 116574. <https://doi.org/10.1016/j.envpol.2021.116574>

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de’ Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., & Schwartz, J. (2019). Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environment International*, *124*, 170‑179. <https://doi.org/10.1016/j.envint.2019.01.016>

Talbi, A., Kerchich, Y., Kerbachi, R., & Boughedaoui, M. (2018). Assessment of annual air pollution levels with PM1, PM2.5, PM10 and associated heavy metals in Algiers, Algeria. *Environmental Pollution*, *232*, 252‑263. <https://doi.org/10.1016/j.envpol.2017.09.041>

Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, *90*, 84‑99. <https://doi.org/10.1016/j.csda.2015.04.009>

Wang, Y., Wang, H., & Zhang, S. (2020). Prediction of daily PM2.5 concentration in China using data-driven ordinary differential equations. *Applied Mathematics and Computation*, *375*, 125088. <https://doi.org/10.1016/j.amc.2020.125088>

Wu, H., Liu, H., & Duan, Z. (2020). PM2.5 concentrations forecasting using a new multi-objective feature selection and ensemble framework. *Atmospheric Pollution Research*, *11*(7), 1187‑1198. <https://doi.org/10.1016/j.apr.2020.04.013>

Xing, H., Wang, G., Liu, C., & Suo, M. (2021). PM2.5 concentration modeling and prediction by using temperature-based deep belief network. *Neural Networks*, *133*, 157‑165. <https://doi.org/10.1016/j.neunet.2020.10.013>

Xu, X., Tong, T., Zhang, W., & Meng, L. (2020). Fine-grained prediction of PM2.5 concentration based on multisource data and deep learning. *Atmospheric Pollution Research*, *11*(10), 1728‑1737. <https://doi.org/10.1016/j.apr.2020.06.032>

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, *10*(7), 373. <https://doi.org/10.3390/atmos10070373>

Zhang, B., Zhang, H., Zhao, G., & Lian, J. (2020). Constructing a PM2.5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. *Environmental Modelling & Software*, *124*, 104600. <https://doi.org/10.1016/j.envsoft.2019.104600>

Zhou, Y., Chang, F.-J., Chang, L.-C., Kao, I.-F., Wang, Y.-S., & Kang, C.-C. (2019). Multi-output support vector machine for regional multi-step-ahead PM2.5 forecasting. *Science of The Total Environment*, *651*, 230‑240. <https://doi.org/10.1016/j.scitotenv.2018.09.111>