ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 9, No. 4, November 2021, pp.1013-1028

Comparison of multiple machine learning algorithms for urban air quality forecasting

Maryam Aljanabi[1], Mohammad Shkoukani[2], Mohammad Hijjawi[3]
[1,2,3] Computer Science Department, Applied Science Private University

ABSTRACT

Environmental air pollution has become one of the major threats to human lives nowadays in developed and developing countries. Due to its importance, there exist various air pollution forecasting models, however, machine learning models proved one of the most efficient methods for prediction. In this paper, we assessed the ability of machine learning techniques to forecast NO_2, SO_2, and PM_{10} in Amman, Jordan. We compared multiple machine learning methods like artificial neural networks, support vector regression, decision tree regression, and extreme gradient boosting. We also investigated the effect of the pollution station and the meteorological station distance on the prediction result as well as explored the most relevant seasonal variables and the most important minimal set of features required for prediction to improve the prediction time. The experiments showed promising results for predicting air pollution in Amman with artificial neural network outperforming the other algorithms and scoring RMSE of 0.949 ppb, 0.451 ppb, and 5.570 $\mu g/m^3$ for NO_2, SO_2, and PM_{10} respectively. Our results indicated that when the meteorological variables were obtained from the same pollution station the results were better. We were also able to reduce the time by reducing the set of variables required for prediction from 11 down to 3 and achieved major time improvement by about 80% for NO_2, 92% for SO_2, and 90% for PM_{10}. The most important variables required for predicting NO_2 were the previous day values of NO_2, humidity and wind direction. While for SO_2 they were the previous day values of SO_2, temperature, and wind direction values of the previous day. Finally, for PM10 they were the previous day values of PM_{10}, humidity, and day of the year.

Keywords: Air pollutants, Machine learning, Supervised learning, Neural networks.

Corresponding Author:

Mohammad Shkoukani
Computer Science Department
Applied Science Private University
Amman, Jordan
m.shkokani@asu.edu.jo

1. Introduction

Due to the increased population on earth, urbanization increased, and with it all sorts of industrialization and transportation. Air pollution refers to the existence of contaminating pollutants in the atmosphere that damages the health of humans [1]. Our atmosphere contains many pollutants from a plethora of areas such as the new chemicals being developed, the combustion of fossil fuels, the heavy usage of transportation systems, heating systems, and much more. This all leads to adverse health effects and increased mortality rates in humans as well as affecting the various species living on earth [2]. The most significant pollutants are ozone (O_3), suspended particle matter (PM), nitrogen oxides (NO_x), carbon monoxide (CO), sulfur dioxide (SO_2), pesticides and other pollutants that are harmful to human's health [3]. In this research, we focused on NO_2, SO_2, and PM_{10}. Suspended particulate matters refer to suspended fine particles in the atmosphere. They may be the result of dust, wind, forest fires or human-made pollution such as industrial processes, car emissions, etc. and can be inhaled and affect the lungs deeply. They are distinguished based on their size with the two main types being PM_{10} and $PM_{2.5}$. PM_{10} are particles with a diameter that is <= 10μm and at the same time > 2.5μm. $PM_{2.5}$ are particles that have a diameter that is <= 2.5 μm [4, 5]. NO_2 is caused when nitrogen oxide is released into the atmosphere. It is caused by natural sources as well as anthropogenic sources such as fossil fuel combustion

resulting from heating systems, power generation, and motors engine emissions [6]. SO_2 pollutant is also caused by natural and man-made sources such as emissions from transportation systems, industry, domestic emissions, power generation emissions, and fuel combustion processes [7-9]. These pollutants are not only harmful to humans but also for the whole ecosystem. Some chemicals that result from human activities cause crops to wither and some emissions have damaged the ozone layer that protects the Earth, this causes more solar radiation to get into the planet's surface which leads to vital skin diseases [10]. The severity of the impact of air pollution led countries to develop indices that are used to assess the quality of the air, whether it's safe for individuals or not [11]. Scientists have been working on forecasting future air pollution levels through the use of statistical models, mathematical simulations such as dispersion models, and chemical and physical equations such as photochemical models. Such models do not use artificial intelligence techniques and instead use pure mathematical and statistical approaches. Since these models have their limitations when it comes to dealing with large datasets, scientists recently started using machine learning techniques for predicting air quality [12-14]. The use of monitoring sensors enabled machine learning scientists to enter the field of air quality forecasting since these sensors are being used to measure air pollutant concentrations and store them in databases. These readings are immensely helpful for machine learning scientists to use them to forecast future levels of air pollution [15]. Machine learning is used in many areas of our lives nowadays and it started being used in the environmental science field in the 1990s. It is used in various environmental areas such as weather forecasting, air quality prediction, ecological modeling, snow, ice and forests monitoring, etc. [16]. Despite their wide application range, machine learning adoption in the environmental science has not been as fast as it is in other areas. Perhaps this is due to the lack of education of machine learning in natural sciences, the absence of communication between machine learning and natural science scientists, or the unavailability of natural data. However, since more data is being collected in the natural world nowadays, the focus on machine learning in the environmental field is growing and is showing promising results as compared to classical statistical methods, because machine learning has better ability to model complex and non-linear relationship between data that exists in the natural world [17]. Multiple machine learning techniques have been used to forecast air pollutants and the results vary from one research to another depending on the dataset at hand, the country of study as well as the pollutant being forecasted [18]. This research focused on forecasting NO_2, SO_2, and PM_{10} in Amman, Jordan, and specifically, in the area of King Al-Hussein Public Parks for one day ahead. The final regression model predicted the numerical concentrations of the four pollutants mentioned earlier. We conducted a comparison between multiple machine learning models which are multi-layer perceptron neural networks (MLP), support vector regression (SVR), extreme gradient boosting (XGB), and decision tree regression (DTR). Then we explored the effect of seasonal variables and which seasonal variables could be used instead of multiple ones to reduce the number of features. A further reduction in features was made in the feature selection step for each of the pollutants mentioned above to reduce the time and cost needed to predict them. We also experimented with different dataset combinations to find the dataset that yielded the best results. This paper has the following structure. The section titled related work provided background information about The use of machine learning techniques to predict air quality alongside researches done in this field that produced promising results. The materials and methods section illustrated several aspects of our research including the dataset, the dataset preprocessing, the feature engineering, the noise removal, the feature selection alongside what performance evaluation metrics were used in this paper. The experimental results and discussion section showed the main results and findings of this research paper, each result was discussed properly and thoroughly. Finally, the conclusion and future work section contained a summary of this research and provided further ideas for researchers who are interested in this field.

2. Related work

Air quality prediction is usually treated as a supervised learning problem when the machine learning algorithm trains on an existing historical dataset containing the input and the desired output to be able to predict future levels of air pollution [19]. Some researchers treated it as a regression problem when they forecasted the numerical concentration of pollutants while others treated it as a classification problem that involves predicting categorical variables, such as high-risk/low-risk, low/medium/high, etc. [20]. Various machine learning algorithms were used in the topic of air quality prediction and many showed great performance as compared to chemical and physical models. Most researches used ANN which is a machine learning algorithm that mimics how neurons in the brain work [21]. This algorithm showed outstanding performance most of the time and was preferred by many researchers as it has many variations and types. A study was conducted to forecast ozone, NO_2, and $PM_{2.5}$ in six Canadian cities in [12]. The author compared multiple variations of ANN and concluded that Online-Sequential Extreme Learning Machine (OS-ELM) outperformed the other methods. In another study

in [22], the authors applied an optimized ANN to predict PM_{10} concentration. The main finding of the study is using stochastic variables analysis to reduce the number of required variables needed for PM_{10} forecasting. Another type of ANN called Cyclic Reservoir with Jumps (CRJ) was used in [23] to predict ozone levels in Croatia in two cities which are Osijek and Kopački. The CRJ was compared to Radial Basis Function (RBF), MLP, Multiple Linear Regression (MLR), and linear regression (LR) and outperformed them all and scoring the lowest errors in Osijek with 91.86 for Mean Square Error (MSE) and 7.134 for Mean Absolute Error (MAE). PM_{10} and $PM_{2.5}$ were predicted in Tehran, Iran in [1] using a mixture of meteorological and seasonal variables. The study compared SVR, Geographically Weighted Regression (GWR), ANN, and Non-linear Autoregressive Exogenous Neural Network (NARX). The study also highlighted the improvement achieved by using a noise eliminating filter named Savitzky-Golay filter. The final results showed that NARX was superior to the other methods used, also the time required for prediction was 14s for PM_{10} and 17s for $PM_{2.5.}$ A study in [24] proposed a model to predict total suspended particles (TSP) and PM_{10} in Salt, Jordan using ANN. The ANN type used in the research was ANNAREX in Matlab and the results showed an MSE of 219.7853 and 1010.7 for PM_{10} and TSP respectively. In [25] long short-term memory neural network extended (LSTME) model was developed to predict hourly $PM_{2.5}$ in Beijing, China. The authors compared spatiotemporal deep learning (STDL), autoregressive moving average (ARMA), the time delay neural network (TDNN), SVR, LSTM, and LSTME. The results indicated the superiority of the developed LSTME with Root Mean Square Error (RMSE) and MAE of 12.60, and 5.46 respectively. SVR is a nonlinear generalization algorithm that generalizes well to new data, it focuses on increasing the margin between boundary points of classes which are also called support vectors and creating a hyperplane that separates them [26]. SVR also showed great results and was preferable to ANN sometimes because it requires fewer parameters for optimization. SVR was implemented in [27] to forecast SO_2, NO_x, nitrogen monoxide (NO), NO_2, CO and respirable suspended particles (RSP) in Hong Kong, China. The SVR was compared to RBF and the result showed that SVR had higher performance. In another study in [28] also in Hong Kong, China, SVR was used to predict CO, NO_2, NO, NO_x, SO_2, $O_{3,}$ and RSP. The comparison was done between online SVR in which data was fed sequentially into the model and normal SVR in which data was provided in batch mode. The online SVR showed better results than normal SVR. Another research predicted air quality index in Beijing, Tianjin, and Shijiazhuang, China using SVR and employing meteorological variables alongside the AQI of the previous day in [29]. The best-developed model for Tianjin displayed 42.78, 6.54, and 4.90 for MSE, RMSE, and MAE respectively. A tree or a decision tree (DT) is a graphical upside-down structure starting at the root and ending with the leaves. A tree is constructed during the training stage and it tries to capture the behaviors of the data through splitting into binary branches, also called binary recursive partitioning. When the decision tree is used for regression purpose it is called regression tree or decision tree regression (DTR) [30]. XGBoost is a tree boosting algorithm that is based on the gradient boosting method. This method is also widely used for a range of applications, such as classification and regression problems. Boosting involves combining multiple models to increase the performance. Gradient boosting is one type of boosting in which the gradient boosting method is used to enhance the tree. XGBoost is being used in many machine learning areas due to using fewer resources and producing good results [31]. These two algorithms are used less than ANN and SVR. XGBoost was used in Tianjin, China to predict $PM_{2.5}$ in [32]. The hourly dataset included features like PM_{10}, O_3, NO_2, SO2, and CO. It covered the period from December 1, 2016, to December 30, 2016. They compared multiple regression models, namely: XGBoost, Random Forest, MLR, DTR, and SVR. The results showed that the model that outperformed the other models was XGBoost with R^2 of 0.9520, RMSE of 17.298, and MAE of 11.774. In [33] the authors predicted $PM_{2.5}$ alongside studying feature importance. The dataset in the study contained daily $PM_{2.5}$ concentrations, climate variables, as well as satellite variables like Aerosol optical depth (AOD), measured at 3 km and 10 km. The researchers implemented Random forest, XGBoost, and deep learning. The results showed that XGBoost produced the best results without AOD at 3 km with R^2 of 0.8 and MAE of 10.0 and RMSE of 13.62. The feature importance study showed that $PM_{2.5}$ lag1 (meaning $PM_{2.5}$ value of the previous day) was the most important in the prediction process. Since choosing the best algorithm highly depends on the dataset and other factors in the prediction process, we compared the algorithms that showed promising results in the previously mentioned papers, namely: ANN, SVR, XGBoost and we also wanted to evaluate the performance of DTR since it was rarely used and since XGBoost is a form of trees.

3. Materials and methods

3.1. The datasets

The location of this study is Amman, which is the capital of Jordan. It is an increasingly expanding city with heavy usage of transportation systems, especially cars and buses [34]. The location of Jordan can be seen in

Figure 1. We obtained the data that we worked on from two sources. The air pollution data, as well as some meteorological data, were obtained from the Jordanian Ministry of Environment from a station located in King Al-Hussein Public Parks (KHP). But since this station has only four meteorological variables, we looked for the closest weather station to obtain more meteorological variables that could be of use. The closest station found was located in the Applied Science Private University (ASU) which is only 9km away from KHP station.
Figure 2 shows these two stations as seen in Google Maps and the distance between them. The red pin shows the location of KHP and the yellow pin is the location of the ASU. The blue line is the distance measured in Google Maps. The King Al-Hussein Public Parks dataset included the daily average concentration of NO_2 (ppb), SO_2 (ppb), and PM_{10} ($\mu g/m^3$) alongside 4 meteorological variables which are temperature (°C), wind speed (km/h), wind direction (°), and relative humidity (%) [35]. The ASU climate dataset contained meteorological variables, namely, air pressure (hpa), wind direction (°), wind speed (km/h), humidity (%), temperature (°C), soil surface temperature (°C), subsoil temperature (°C), precipitation (mm), direct radiation (W/m²) and dew point temperature (°C) [36].

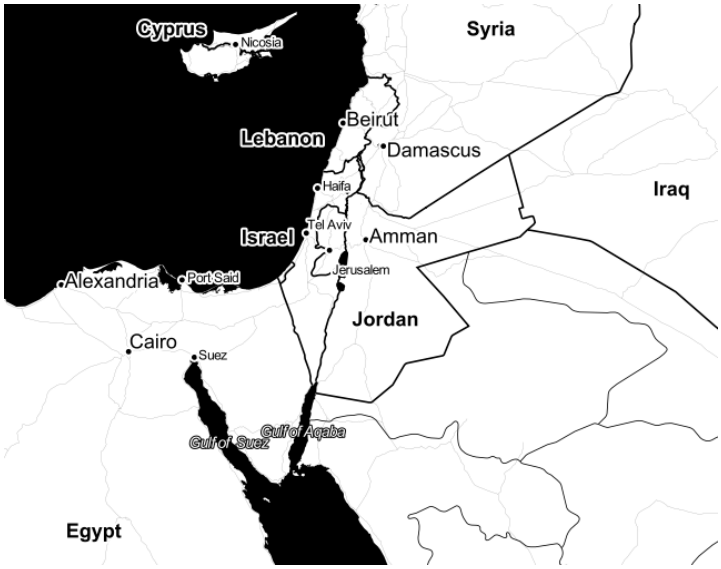

Figure 1. Location of Jordan [37]

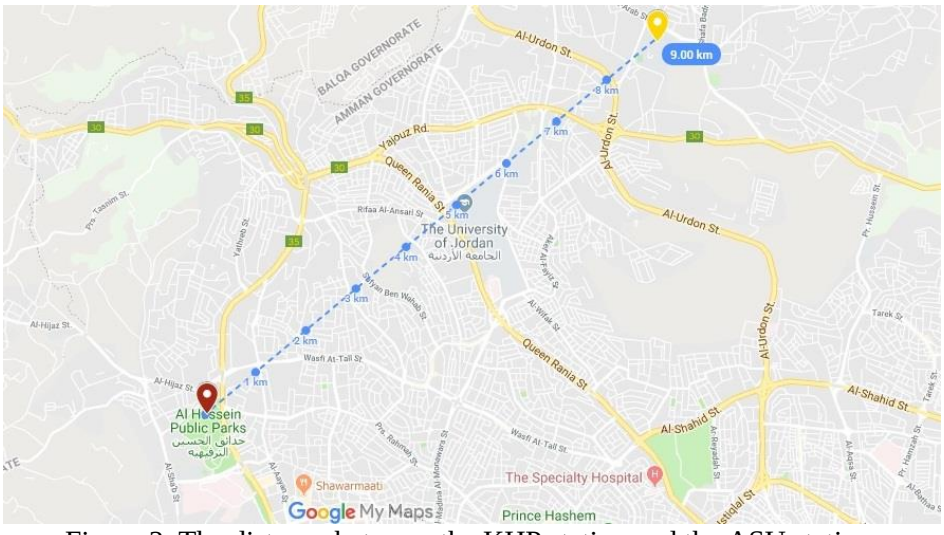

Figure 2. The distance between the KHP station and the ASU station.

We aggregated three combinations of these datasets. The first, which we will call dataset 1, contained the features from KHP station only. The second, called dataset 2, contained ASU station's meteorological data combined with KHP station's pollution data only. The third, called dataset 3, consisted of KHP station's pollution and meteorological data combined with the remaining meteorological data from ASU station. The reason for these dataset combinations is to find the combination that can achieve the highest performance for air quality prediction. We wanted to check if the additional meteorological variables from the ASU station

would enhance the prediction results or not. Moreover, we wanted to find the effect of taking the meteorological variables from a station far from the pollution station. Table 1 illustrates the datasets, the stations, the sources and the features in each dataset combination.

Table 1. Datasets descriptions

	Source	Total Features	Total	Time	Records
1	KHP *dataset 1*	O_3, PM_{10}, NO_2, SO_2, Date, Temperature, Humidity, Wind Speed, Wind Direction.	9	01-05-2014 to 04-06-2019	1886
2	KHP (pollutants) + ASU (meteorological) *dataset 2*	O_3, PM_{10}, NO_2, SO_2, Date, Temperature, Wind Speed, Wind Direction, Humidity, Soil Surface, Subsoil Temperature, Temperature, Precipitation, Dew point, Air Pressure, Direct Radiation.	16	14-05-2015 to 18-02-2019	1376
3	KHP (pollutants and meteorological) + ASU (remaining meteorological) *dataset 3*	O_3, PM_{10}, NO_2, SO_2, Date, Temperature (KHP), Humidity (KHP), Wind Speed (KHP), Wind Direction (KHP), Subsoil Temperature (ASU), Soil Surface Temperature (ASU), Precipitation (ASU), Dew point (ASU), Air Pressure (ASU), Direct Radiation (ASU).	16	14-05-2015 to 18-02-2019	1376

3.2. Data preprocessing

Our datasets contain a total of 161, 323, 289 missing values for dataset 1, 2 and 3 respectively. So dataset 1 has the least amount of missing values. Since our dataset is a time-series dataset, we cannot remove the missing values because we cannot simply remove days from the dataset time-line. There are many methods for filling out missing values. We used the interpolation method to treat the values that are missing in our time-series dataset which is using a mathematical function to substitute the missing values in the dataset. Since the interpolation method cannot fill the missing data that appears at the beginning of the dataset accurately, we removed the first month of dataset 1 since it has a lot of missing values at the beginning. At this point, dataset 1 interval changed to cover the period from June 4, 2014, to June 4, 2019, with 1826 records which is 5 years of daily data.

3.3. Feature engineering

This step is crucial in the case of time-series data. It means adding more meaningful features to our dataset which may help in the prediction process. These additional features will be added to each of the 3 datasets we have. Since the machine learning algorithm cannot deal with a "Date" field, so in our case, we extracted the important features from the "Date" field and stored them in multiple features. The date variables, also called seasonal variables, that we extracted are the day of the week, the day of the month, the day of the year, the month, the special day (whether a day is a holiday or a weekend or not), and the season (winter, spring, summer, and autumn). Seasonal variables can influence the behavior of pollutants, hence the importance of adding them. Table 2 shows the seasonal features added for each of the three datasets.

Table 2. Added seasonal variables

Features	Values
Month	1: January to 12: December
Day of the week	0: Monday to 6: Sunday
Day of the month	1 to 31
Day of the year	1 to 365
Season	1: Winter, 2: Spring, 3: Summer, 4: Autumn
Special Day	1: special day, 0: not a special day

3.4. Noise removal

Time-series data tend to contain a lot of noise which makes it harder for the machine learning algorithms to learn from them and make accurate predictions. The noise removal stage in time-series is one of the most important stages since it can prepare the dataset properly for the machine learning algorithm and eliminate the noise without losing important information in the data. The importance of using the noise removal filter was highlighted in [1] where the authors discovered an immense improvement after using the filter. The importance of using smoothing filters was also mentioned in several studies concerning time-series smoothing [38, 39]. There are various denoising filters that could be used, one of the most powerful filters is the Savitzky-Golay filter. We tried different values for the parameters of the filter and arrived at the best combination which was a window length of 25 and a polynomial of 4. This configuration made the data smoother while preserving the peaks and the important information, thus no data loss was encountered. The filter was applied to all the numerical features in the dataset. Figures 3, 4, and 5 illustrate the effect of applying the filter to the data of NO_2, SO_2, and PM_{10} respectively. The lighter line indicates the original unfiltered data while the darker line is the filtered data. The smoothing filter also removed the outliers which are the extreme values in the dataset and smoothed them. After applying the noise removal filter, the method used to normalize the data was the MinMax scaler which transformed the values into a unified range between 0 and 1 so that they have the same weight when the machine learning algorithms train on them.

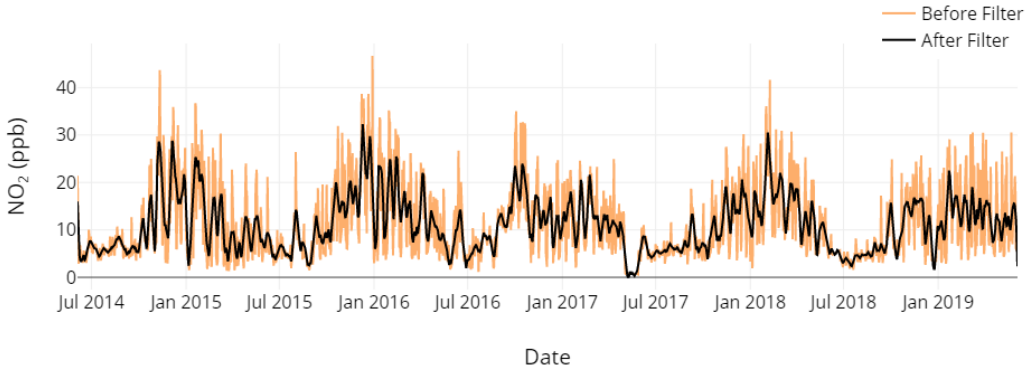

Figure 3. NO_2 data before and after filtering

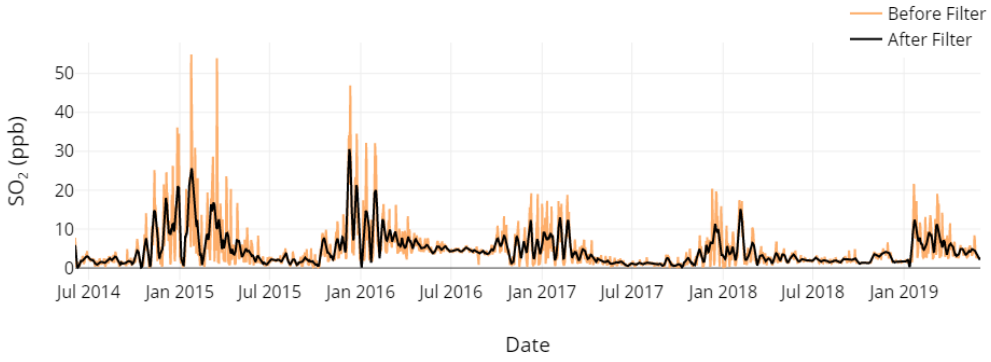

Figure 4. SO_2 data before and after filtering

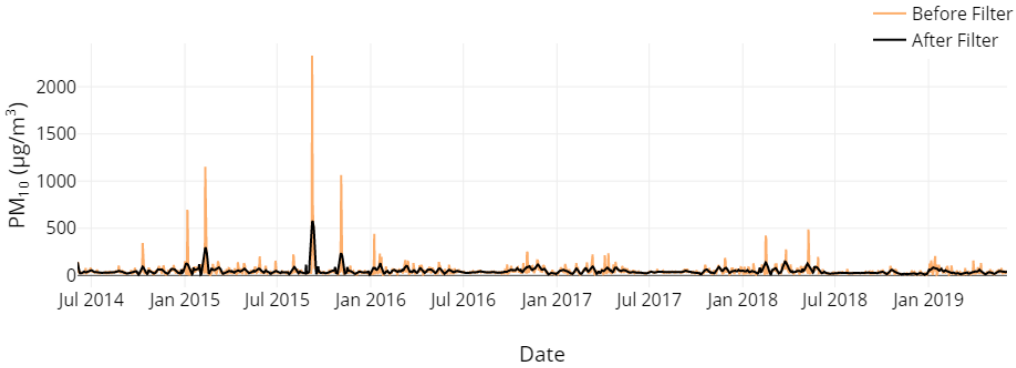

Figure 5. PM_{10} data before and after filtering

3.5. Feature selection

The feature selection step is crucial while building a predictive model in machine learning because it can greatly decrease the computational power and time taken for prediction, as well as improve the accuracy. This step focuses on selecting only a subset of the features used as input for the model, it chooses the most important features for the prediction model and gets rid of the irrelevant ones. The main techniques of feature selection are the filter and the wrapper methods [40]. The filter method uses a filtering algorithm in order to find the most effective features corresponding to the output that we want to predict. For example, the Pearson correlation-based filter depends on using the correlation between each input feature and the output of the model, it's a measure of how related these variables are. The wrapper method, unlike the filter approach that is generic and doesn't depend on any model, the wrapper is rather model dependent. It works by finding the best subset of features that scores the best result using a specific model that is specified by the researcher. There are many types of wrappers that differ on the basis of how they find the best subset of features. For example, the forward wrapper starts adding features to an empty set one by one. In each phase, the feature subset that yields the best result when used in the model is kept and the others are discarded. This approach is more comprehensive and may outperforms the filter since it is concerned with subsets of features rather than individual features relationships with the output, yet it can be computationally expensive especially for large datasets [41]. In our work, we used the forward wrapper method to perform the feature selection stage. The most significant features that influence the prediction of each pollutant differ and depend on the dataset used and its location. In a study conducted to reveal the most influential variables on ground-level ozone in Eastern Texas, USA [42], it was found that NO_2 alongside wind speed, and wind direction had the greatest influence, while temperature did not play a vital role in increasing ozone. However, in other studies, it was shown that temperature and humidity highly influenced ozone concentrations [43]. An EPA environmental report also indicated the importance of temperature, humidity and wind speed on ozone levels [44]. In [45] it was found that most pollutants decrease with the increase of humidity in Dhaka, Bangladesh. Temperature, humidity and precipitation were found dominant for PM_{10} concentration in Andean, Colombia [46], while wind gust was the most important factor in Switzerland as well as precipitation and seasonal variables [47]. For NO_2, some experiments showed the importance of wind speed on its production in [48]. On the other hand, in another study, the wind direction was found to have the highest impact on NO_2 concentration while wind speed was found of little importance in Gothenburg, in south-west Sweden [49]. This shows how complex is the problem of uncovering the most important variables affecting a certain pollutant. This variation could be due to a plethora of aspects such as the location of the station of the dataset, its elevation from sea-level, the distance of the dataset from crowded streets or factories, the time period of the dataset, the seasons it covered, the climate of the country of the dataset and more [50].

3.6. Performance evaluation metrics

In order to measure the performance and compare the results of the different models used in our experiments, we used the Coefficient of Determination (R^2), the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE) as the performance evaluation metrics which are specifically used for regression models. In all the following equations, N stands for the number of samples, P is the predicted value, and A is the actual value [1, 12].

$$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}[(P_i - \bar{P})(A_i - \bar{A}]}{\sigma_P \sigma_A}\right]^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - A_i)^2} \quad (2)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - A_i| \quad (3)$$

4. Experimental results and discussion

The experiments in this research were done using python 3. The experiments were carried out using HP laptop with Windows 8.1 64-bit, a Core-i5 2.2 GHz processor and 4GB RAM.

4.1. Model and dataset selection

The first step in the experiments is applying the four algorithms we are comparing to all three datasets with the three pollutants. For each pollutant prediction model, the input to the model is the pollutant itself from the previous day alongside the previous day seasonal variables and meteorological variables. The output is the pollutant concentration of the next day. The model and the dataset that will score the highest will be selected

for the next step which involves reducing the number of features. Table 3 illustrates the results obtained for all the NO_2 models. Dataset 1 achieved the highest overall results with a small difference from the other datasets. The neural network gave the maximum values for each dataset, with the top result being 96.160%, 1.005 ppb and 0.757 ppb for R^2, RMSE, and MAE respectively. As seen in Table 4, the SO_2 results are fairly close to the NO_2 results. Not only the best model is the neural network and the best dataset is dataset 1, but also the experimental values are similar to the ones produced by NO_2. The top result showed R^2 of 96.095%, RMSE of 0.498 ppb, and the MAE was 0.338 ppb. As for the final pollutant, PM_{10}, Table 5 describes the outputs. The highest result achieved was 91.281%, 6.759 $\mu g/m^3$, and 4.783 $\mu g/m^3$ for R^2, RMSE, and MAE respectively. The best model that provided these results was the neural network of dataset 1.

Table 3. NO_2 results

Dataset	Model	R^2 (%)	RMSE (ppb)	MAE (ppb)
Dataset 1	MLP	96.160	1.005	0.757
	SVR	91.911	1.459	1.194
	XGBoost	92.200	1.437	1.191
	DTR	86.551	1.881	1.472
Dataset 2	MLP	95.163	1.190	0.931
	SVR	90.456	1.671	1.287
	XGBoost	91.791	1.550	1.160
	DTR	78.928	2.483	1.752
Dataset 3	MLP	95.619	1.132	0.873
	SVR	91.082	1.616	1.185
	XGBoost	91.805	1.549	1.155
	DTR	83.960	2.167	1.631

Table 4. SO_2 results

Dataset	Model	R^2 (%)	RMSE (ppb)	MAE (ppb)
Dataset 1	MLP	96.095	0.498	0.338
	SVR	93.805	0.627	0.464
	XGBoost	93.842	0.625	0.489
	DTR	86.321	0.931	0.531
Dataset 2	MLP	94.550	0.580	0.378
	SVR	92.657	0.674	0.468
	XGBoost	92.681	0.673	0.509
	DTR	84.382	0.983	0.640
Dataset 3	MLP	94.898	0.562	0.398
	SVR	92.889	0.663	0.468
	XGBoost	92.908	0.662	0.498
	DTR	85.101	0.960	0.683

Table 5. PM_{10} results

Dataset	Model	R^2 (%)	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)
Dataset 1	MLP	91.281	6.759	4.783
	SVR	90.570	7.029	4.686
	XGBoost	89.033	7.580	5.302
	DTR	70.322	12.469	8.267
Dataset 2	MLP	89.816	8.064	6.211
	SVR	86.080	9.428	6.849
	XGBoost	83.824	10.163	8.353
	DTR	65.185	14.910	10.204

Dataset	Model	R^2 (%)	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)
Dataset 3	MLP	90.834	7.650	5.516
	SVR	87.886	8.795	6.437
	XGBoost	84.716	9.879	6.473
	DTR	67.341	14.441	9.321

From the above Tables, we can see that MLP, SVR, and XGBoost results were fairly similar with MLP being in the lead with a small difference and SVR and XGBoost performing very similarly to each other. This shows that all of the three mentioned algorithms performed well in our datasets and were able to detect the patterns and predict the pollution concentration with great performance. Although it's fast, yet DTR, on the other hand, had the worst performance and it was always the lowest for all the pollutants. The large difference between results of different datasets using DTR for the same pollutant is due to its instability, meaning the result differs a lot when a small change in the dataset occurs. Dataset 1 proved to be the most reliable for all pollutants. Its results were the best. However, the results of the other datasets were close as well, especially for MLP. Dataset 3 showed better results than dataset 2, this could be because dataset 2 has the meteorological variables taken from KHP, which is the same station for the pollutant variables. This shows that the results are the best when the meteorological variables are taken from the same station that measures pollutants. Yet, if there was no meteorological station at the same place, then the results would not worsen so much if the two stations were not too far away from each other. In our case, there was a difference of only 9km between the two stations, and this was reflected in a slight decrease in the performance of the models. Another remark on the datasets is that the additional meteorological variables from the ASU weather station were irrelevant and did not improve the prediction results. Predicting NO_2 and SO_2 scored higher results than PM_{10}. Their results were fairly close since both pollutants are produced by similar conditions and we can even notice that they have similar patterns. PM_{10} had the lowest prediction result compared to the other two since this pollutant is affected by unpredictable weather conditions like dust storms as well as other factors. Yet its results are still quite good and promising. Overall all the experiments showed promising results and low error rates. The final result of this step is choosing dataset 1 and MLP ANN as the best model and it will be used to work with the next steps.

4.2. Seasonal variables feature importance

This step involves studying the most relevant seasonal variables and discarding the rest. Since we already have the day of the year variable, the algorithm may already be able to conclude the month, season, the day of the month, and the day of the week variables from the day of the year. For this reason, we performed two experiments to help understand the importance of the day of year feature, one experiment was conducted with all the seasonal variables except the day of the year, and another experiment was performed with only the day of the year alongside the special day feature, since this one cannot be concluded from the day of the year and it varies depending on the holidays that may change from year to year. The features included in the experiment which yielded the best result were chosen for this step. Tables 6,7, and 8 show the results for NO_2, SO_2, and PM_{10} respectively. As shown in the tables, there is no vast difference between the results, but using the day of the year without the other seasonal variables always yielded the top result. Most results contained a difference of nearly 0.1% except for PM_{10} that has a difference of about 2%. Yet the major difference lies in time, there is a visible improvement in time when using the day of the year alone without the other variables, clearly because the number of features has been reduced. This clearly shows that the month, the day of the week, the day of the month, and the season do not contribute to the prediction system and they are unnecessary. Since using the day of the year with the special day features instead of the remaining seasonal variables showed an improvement in time and performance, then the output from this step is neglecting the remaining seasonal variables that proved irrelevant.

Table 6. NO_2 seasonal variables comparison

Input Features	**R^2 (%)**	**RMSE (ppb)**	**MAE (ppb)**	**Time (ms)**
Previous day values of NO_2, meteorological variables, day of the year, and special day	96.437	0.968	0.716	226
Previous day values of NO_2, meteorological	96.103	1.013	0.758	655

Input Features	R^2 (%)	RMSE (ppb)	MAE (ppb)	Time (ms)
variables, day of the week, day of the month, month, season, and special day				

Table 7. SO_2 seasonal variables comparison

Input Features	R^2 (%)	RMSE (ppb)	MAE (ppb)	Time (ms)
Previous day values of SO_2, meteorological variables, day of the year, and special day	96.191	0.491	0.330	301
Previous day values of SO_2, meteorological variables, month, season, day of the week, day of the month, and special day	96.008	0.503	0.351	503

Table 8. PM_{10} seasonal variables comparison

Input Features	R^2 (%)	RMSE (µg/m^3)	MAE (µg/m^3)	Time (ms)
Previous day values of PM_{10}, meteorological variables, day of the year, and special day	92.152	6.412	4.432	552
Previous day values of PM_{10}, meteorological variables, day of the week, day of the month, month, season, and special day	90.253	7.14588	5.134	725

4.3. Feature selection results

At this point, we have seven input variables for each pollutant model which are: the pollutant value of the previous day, the meteorological and seasonal variables of the previous day, namely: the humidity, temperature, the day of the year and the special day, wind direction, and wind speed. In this stage, we used the wrapper method to decrease the number of features to the minimum amount possible to improve the performance and decrease time. The following subsections demonstrate how feature selection affected each of the pollutant results. Note that the experiments were carried out using dataset 1 and using the MLP model. For NO_2, we can see in Table 8 that the best subset of features found was NO_2, humidity and wind direction of the previous day. Since NO_2 is generated by emissions and mainly peaks in cold weather, we deduce that wind direction and humidity would impact its production the most. We encountered a great improvement in time of about 80%, while R^2 improved by about 0.1%, the RMSE and MAE decreased to 0.950 ppb and 0.701 ppb respectively. Table 9 shows the results obtained for SO_2. The optimal set of features was SO_2, humidity and wind direction of the previous day, which is also similar to NO_2 optimal subset of features. The time improvement was 92%, while R^2 increased by 0.6% and RMSE, and MAE dropped to 0.491 and 0.330 respectively. Finally, for PM_{10}, we can observe in Table 10 that the best subset of features found was the previous day values of PM_{10}, the humidity, and the day of the year, with a time improvement of around 90%. The important features of PM_{10} can also tell us how this pollutant is highly influenced by the time of the year and also by weather conditions. The increase in R^2 was nearly 2% while MAE decreased by more than 1% and RMSE dropped to 5.570. The previous results indicate that feature selection improved the results of all pollutants in various degrees as well as help us understand the nature of pollutants more and what influences them the most. The R^2, RMSE, and MAE have all been improved, although the greatest improvement was seen in PM_{10}. Another major enhancement was the time. We can see a vast improvement in time from before and after the feature selection as it improved by 80%, 92%, and 90% for NO_2, SO_2, and PM_{10} respectively.

Table 9. NO_2 feature selection results

Features	R^2 (%)	RMSE (ppb)	MAE (ppb)	Time (ms)
Previous day values of NO_2, meteorological variables, day of the year, and special day	96.437	0.968	0.716	226
Previous day values of NO_2, wind direction, and humidity.	96.574	0.950	0.701	45

Table 10. SO_2 feature selection results

Features	R^2 (%)	RMSE (ppb)	MAE (ppb)	Time (ms)
Previous day values of SO_2, meteorological variables, day of the year, and special day	96.191	0.491	0.330	301
Previous day values of SO_2, temperature, and wind direction	96.792	0.451	0.291	23

Table 11. PM_{10} feature selection results

Features	R^2 (%)	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	Time (ms)
Previous day values of PM_{10}, meteorological variables, day of the year, and special day	92.152	6.412	4.432	552
Previous day values of PM_{10}, day of the year, and humidity.	94.079	5.570	3.594	52

At this point we have arrived at the optimal results and number of features for every pollutant. The configuration of the machine learning models has all been found through trying different parameter values. Figures 6,7, and 8 illustrate the optimal neural network configurations for each model of the three pollutant models we have. We arrived at this combination of neurons based on experimenting with different neural network configurations and chose the ones that yielded the best performance for each pollutant.

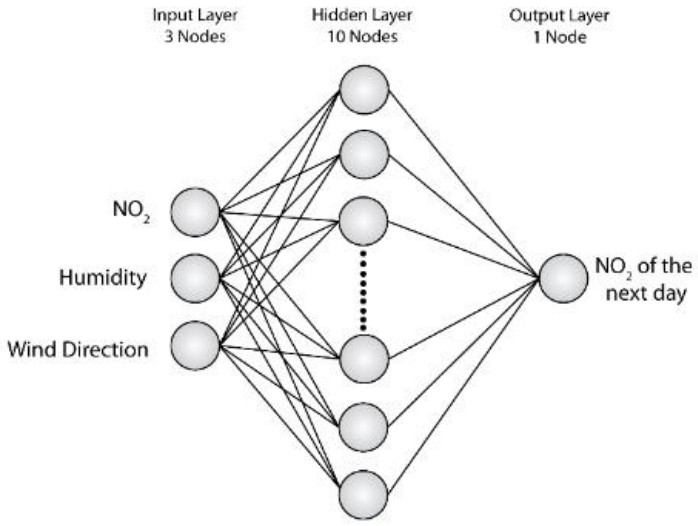

Figure 6. NO_2 model

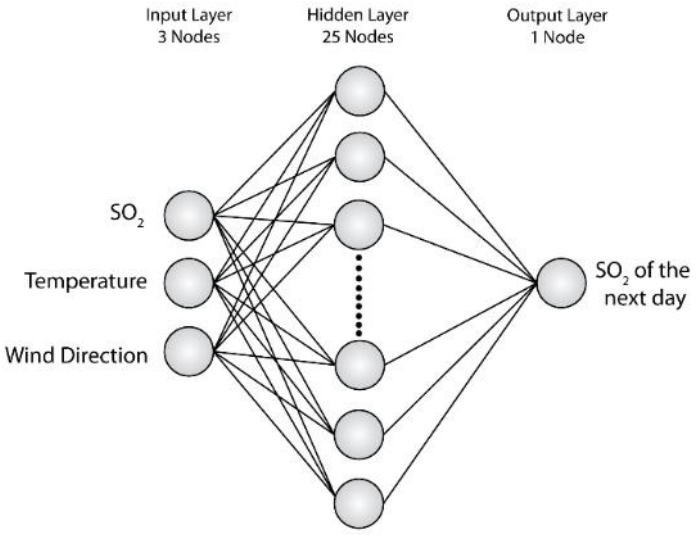

Figure 7. SO_2 model

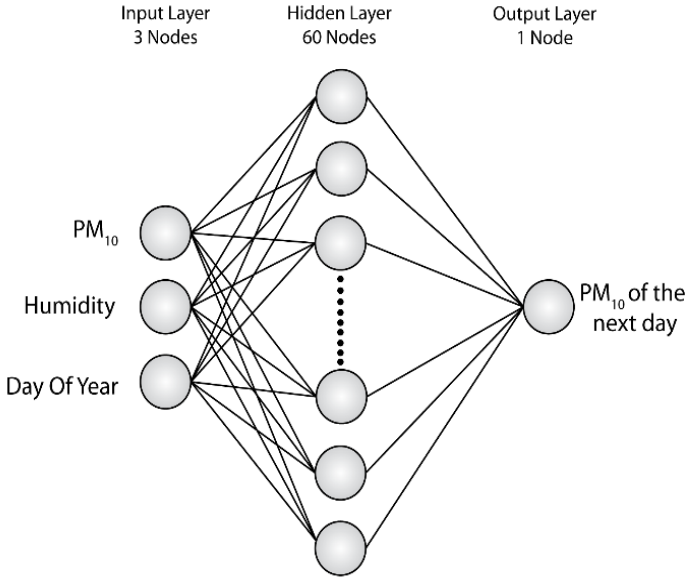

Figure 8. PM_{10} model

Figures 9, 10, and 11 reflect the performance of the final models for NO_2, SO_2, and PM_{10}. We can clearly observe that the predicted samples in the dotted black line fit the actual samples in the light orange, and there is very little error rates.

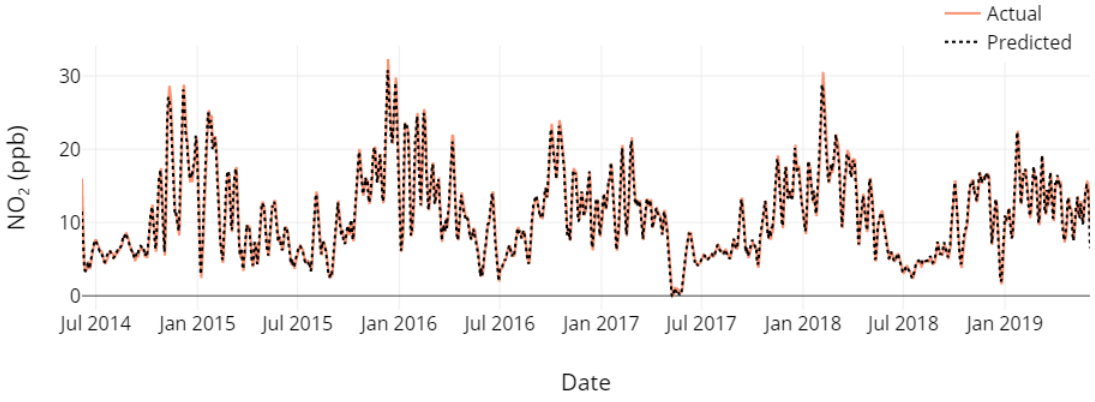

Figure 9. NO_2 time-series actual and predicted

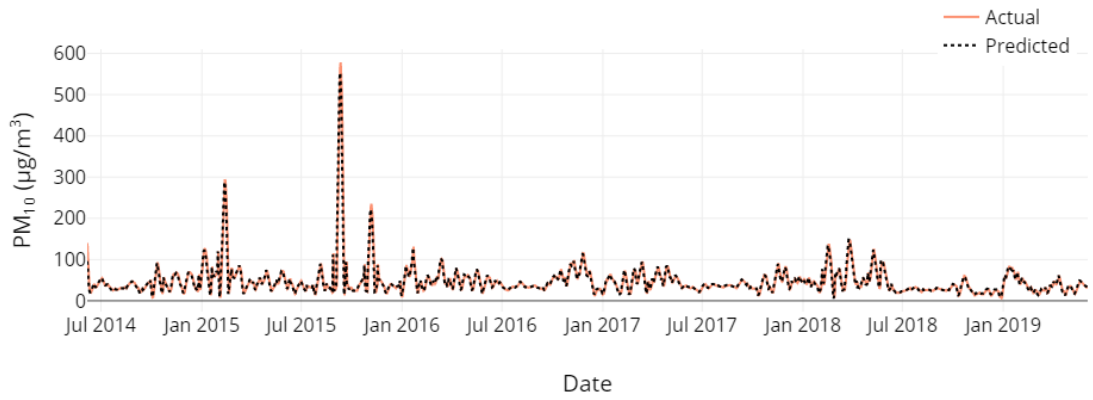

Figure 10. SO_2 time-series actual and predicted

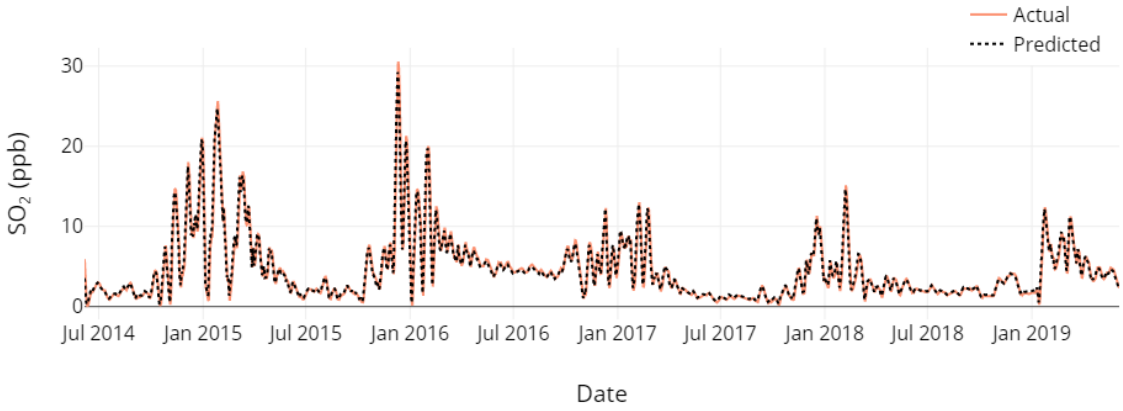

Figure 11. PM_{10} time-series actual and predicted

5. Conclusion and future work

In this research, we built a model to predict air pollution for one day ahead in Amman, Jordan, for four pollutants, namely: NO_2, SO_2, and PM_{10}. the main findings of this research are as follows:

- We worked with three combinations of datasets to uncover the location's importance of the dataset as well as the relevance of some additional meteorological variables to the prediction process. Dataset 1, in which the meteorological and pollution variables were obtained from the KHP station. Dataset 2 in which the meteorological variables were taken from another station which is the ASU station, 9km away from the KHP station, and dataset 3 in which some meteorological variables were taken from KHP and the rest from ASU station. We found the dataset 1 scored the best results, yet the other datasets still performed well too but less than dataset 1. This leads to the conclusion that the prediction is the most accurate when the meteorological station is the same as the pollution station or as close as possible. Another remark on this point is that the additional meteorological variables obtained from the ASU station were irrelevant.
- A comprehensive comparison between MLP ANN, SVR, XGBoost, and DTR was carried out for all the pollutants and all the datasets. The model that outperformed the others was always MLP in the case of all stations and all the pollutants. SVR and XGBoost performed well too especially for dataset 1, but they were slightly less than the performance of MLP. DTR performed poorly compared to the other models and was unstable when the dataset changed.
- A study of seasonal variables importance was carried out which showed that using the day of the year feature instead of the day of the week, day of the month, month, and season generated better results and reduced the time.
- The crucial features for predicting each of the four pollutants were discovered through the feature selection step. All the performance evaluation metrics were improved with major enhancement in time for all pollutants.
- This research achieved a reduction of features for each pollutant model from 11 down to 3 which greatly reduced the time by 80%, 92%, and 90% for NO_2, SO_2, and PM_{10} respectively.

- Machine learning models, especially MLP, showed promising results in the field of air quality prediction with reduced errors and reliable forecasts.
- We built a model for predicting air pollution concentration in Amman, Jordan for the next day, which is the first to be done in Amman using these datasets we worked with. The final model for NO_2 achieved R^2 of 96.574%, RMSE of 0.950 ppb and MAE of 0.701 ppb. Similarly, the SO_2 model scored 96.792%, 0.451 ppb and 0.291 ppb for R^2, RMSE, and MAE respectively. Finally, PM_{10} achieved R^2 of 94.079%, RMSE of 5.570 $\mu g/m^3$, and MAE of 3.594 $\mu g/m^3$.

For future work in this area, it would be great if this model would be applied to online generating data, in which the data readings are fed into the model daily so that it would be possible to continuously predict the pollution levels of the next day. A website or a mobile application could be built if such data and permission from the data owners would be obtained. Ideally, there should be various air pollution and meteorological stations across Amman to allow continuous prediction of air pollution for multiple areas. If they became available in the future, this model could be applied to them with some modifications. If more than one air pollution station was available, it would be possible to add some spatial parameters like the location of the station and its elevation from sea-level. Also, consider adding some meaningful parameters related to pollution like traffic parameters such as the number of passing cars in a day, which we considered but weren't able to obtain in our research.

Acknowledgment

The authors are thankful for the Jordanian Ministry of Environment for granting them permission to use the King Al-Hussein Public Parks dataset collected by the ministry that contains the meteorological and air pollution data used in this research. The authors are also grateful to the Applied Science Private University, Amman, Jordan, for supporting this research fully and for providing the meteorological dataset obtained from the ASU weather station provided by the Renewable Energy Center.

6. References

[1] M. Delavar, A. Gholami, G. Shiran, Y. Rashidi, G. Nakhaeizadeh, K. Fedra, and S. Afshar, "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran," *ISPRS International Journal of Geo-Information,* journal article vol. 8, no. 2, pp. 99-119, 2019.

[2] J. Seinfeld and S. Pandis, *Atmospheric Chemistry And Physics From Air Pollution to Climate Change*, Third ed. New Jersey, USA: John Wiley & Sons, Inc., 2016.

[3] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," *Big Data and Cognitive Computing,* vol. 2, no. 1, pp. 1-15, 2018, Art. no. 5.

[4] United States Environmental Protection Agency (EPA), "Environments and Contaminants, Criteria Air Pollutants," in *America's Children and the Environment*Third ed. Washington, D.C, United States, 2015.

[5] X. Li, X. Chen, X. Yuan, G. Zeng, T. León, J. Liang, G. Chen, and X. Yuan, "Characteristics of Particulate Pollution (PM2.5 and PM10) and Their Spacescale-Dependent Relationships with Meteorological Elements in China," *Sustainability,* vol. 9, no. 12, p. 2330, 2017.

[6] World Health Organization (WHO), *Air quality guidelines for Europe*, Second ed. Copenhagen : WHO Regional Office for Europe, 2000.

[7] Z. Lu, D. G. Streets, Q. Zhang, S. Wang, G. R. Carmichael, Y. F. Cheng, C. Wei, M. Chin, T. Diehl, and Q. Tan, "Sulfur dioxide emissions in China and sulfur trends in East Asia since 2000," *Atmospheric Chemistry & Physics*, vol. 10, pp. 6311-6331, 2010.

[8] I. A. Aljazaery, H. Alhasan, F. N. Al Hachami, and H. T. H. S. Alrikabi, "Simulation study to calculate the vibration energy of two molecules of hydrogen chloride and carbon oxide," *Journal of Green Engineering,* Article vol. 10, no. 9, pp. 5989-6010, 2020.

[9] M. Mahmuddin and A. G. M. Al-dawoodi, "Experimental study of variation local search mechanism for bee algorithm feature selection," *Journal of Telecommunication, Electronic Computer Engineering,* vol. 9, no. 2-2, pp. 103-107, 2017.

[10] L. Bai, J. Wang, X. Ma, and H. Lu, "Air Pollution Forecasts: An Overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, pp. 780-824, 2018.

[11] A. Plaia and M. Ruggieri, "Air quality indices: a review," *Reviews in Environmental Science and Bio/Technology,* vol. 10, no. 2, pp. 165–179, 2011.

[12] H. Peng, "Air Quality Prediction by Machine Learning Methods," Master Thesis, The Faculty of Graduate and Post doctoral Studies (Atmospheric Science), The University of British Columbia, Vancouver, Canada, 2015.

[13] A. M. Abass, O. S. Hassan, A. Rikabi, H. T. Salim, and A. Ahmed, "Potentiometric determination of fexofenadinehydrochloride drug by fabrication of liquid membrane electrodes," *Egyptian Journal of Chemistry*, vol. 64, no. 11, pp. 5-6, 2021.

[14] A. G. M. Al-Dawoodi, "An improved Bees algorithm local search mechanism for numerical dataset," *Universiti Utara Malaysia*, 2015.

[15] C. Bellinger, M. Shazan, M. Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health*, vol. 17, no. 1, pp. 907-926, 2017.

[16] W. Hsieh, *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*. Cambridge, UK: Cambridge University Press, 2009, p. 364.

[17] G. Humphries, D. Magness, and F. Huettmann, *Machine Learning for Ecology and Sustainable Natural Resource Management*. New York, USA: Springer International Publishing, 2018.

[18] Y. Rybarczyk and R. Zalakeviciute, "Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review," *Applied Sciences*, vol. 8, no. 12, pp. 2570-2598, 2018.

[19] S. Marsland, *Machine Learning: An Algorithmic Perspective*, Second ed. London, UK: Chapman & Hall/CRC, 2014, p. 406.

[20] E. Alpaydin, *Introduction to Machine Learning*, Second ed. Cambridge, Massachusetts, USA: The MIT Press, 2010, p. 584.

[21] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Third ed. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2007.

[22] A. Russo, F. Raischel, and P. G. Lind, "Air quality prediction using optimal neural networks with stochastic variables," *Atmospheric Environment*, vol. 79, pp. 822-830, 2013/11/01/ 2013.

[23] A. Sheta, H. Faris, A. Rodan, E. Kovač-Andrić, and A. Al-Zoubi, "Cycle reservoir with regular jumps for forecasting ozone concentrations: two real cases from the east of Croatia," *Air Quality, Atmosphere & Health*, vol. 11, no. 5, pp. 559–569, 2018.

[24] M. Alkasassbeh, A. Sheta, H. Faris, and H. Turabieh, "Prediction of PM10 and TSP Air Pollution Parameters Using Artificial Neural Network Autoregressive, External Input Models: A Case Study in Salt, Jordan," *Middle East Journal of Scientific Research*, vol. 14, no. 7, pp. 999-1009, 2013.

[25] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation," *Environmental Pollution*, vol. 231, no. 1, pp. 997-1004, 2017/12/01/ 2017.

[26] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[27] W. Lu, W. Wang, A. Leung, S. Lo, R. Yuen, Z. Xu, and H. Fan, "Air pollutant parameter forecasting using support vector machines," in *International Joint Conference on Neural Networks*, Honolulu, HI, USA, 2002, vol. 1, pp. 630-635: IEEE.

[28] W. Wang, C. Men, and W. Lu, "Online prediction model based on support vector machine," *Neurocomputing*, vol. 71, no. 4-6, pp. 550-558, 2008.

[29] B. Liu, A. Binaykia, P. Chang, M. Tiwari, and C. Tsao, "Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang," *PLoS ONE*, vol. 12, no. 7, pp. 1-17, 2017.

[30] G. Moisen, "Classification and regression trees," in *Encyclopedia of Ecology*First ed. Oxford, U.K: Elsevier, 2008.

[31] B. Zhai and J. Chen, "Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China," *Science of the Total Environment*, vol. 635, no. 1, pp. 644–658, 2018.

[32] B. Pan, "Application of XGBoost algorithm in hourly PM2.5 concentration prediction," in *IOP Conference Series Earth and Environmental Science*, Harbin, China, 2018, vol. 113, pp. 12127-12135: IOP.

[33] M. Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data," *Atmosphere*, vol. 10, no. 7, pp. 373-392, 2019.

[34] R. Potter, K. Darmame, N. Barham, and S. Nortcliff, "An Introduction to the Urban Geography of Amman, Jordan," in *Reading Geographical Paper*, vol. 182First ed. Reading, UK The University of Reading, 2007.
[35] Daily Pollution Concentrations in King Al-Hussein Public Parks Station Dataset, 2019.
[36] ASU Weather Station Dataset, 2019.
[37] Stamen and OpenStreetMap. (10-Nov-2019). *Stamen Maps*. Available: http://maps.stamen.com/toner/#6/31.588/35.552
[38] R. Alrumaih and M. Al-Fawzan, "Time Series Forecasting Using Wavelet Denoising an Application to Saudi Stock Index," *Journal of King Saud University - Engineering Sciences,* vol. 14, no. 2, pp. 221-233, 2002.
[39] A. DeLivera, R. Hyndman, and R. Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing," *Journal of the American Statistical Association,* vol. 106, no. 496, pp. 1513-1527, 2011.
[40] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proceedings, Twentieth International Conference on Machine Learning*, Washington, DC, United States, 2003, vol. 2, pp. 856-863.
[41] G. Naqvi, "A Hybrid Filter-Wrapper Approach for Feature Selection," Master Thesis, Department of Technology, Örebro University, Örebro, Sweden, 2012.
[42] A. Gorai, F. Tuluri, P. Tchounwou, and S. Ambinakudige, "Influence of local meteorology and NO2 conditions on ground-level ozone concentrations in the eastern part of Texas, USA," *Air Quality, Atmosphere & Health,* vol. 8, no. 1, pp. 81–96, 2015.
[43] V. Valuntait, V. Šerevicien, R. Girgždien, and D. Paliulis, "Relative Humidity and Temperature Impact to Ozone and Nitrogen Oxides Removal Rate In The Experimental Chamber," *Journal Of Environmental Engineering and Landscape Management,* vol. 20, no. 1, pp. 35–41, 2012.
[44] United States Environmental Protection Agency (EPA). (2017, 20-Nov-2019). *Ambient Ozone*. Available: https://www.epa.gov/roe
[45] I. Kayes, S. Shahriar, K. Hasan, M. Akhter, M. Kabir, and M. Salam, "The relationships between meteorological parameters and air pollutants in an urban environment," *Global Journal of Environmental Science and Management,* vol. 5, no. 3, pp. 265-278, 2019.
[46] M. González-Duque, J. Cortés-Araujo, and H. Aristizábal-Zuluaga, "Influence of meteorology and source variation on airborne Pm10 levels in a high relief tropical Andean city," *Revista Facultad de Ingeniería Universidad de Antioquia,* no. 74, pp. 200-212, 2015.
[47] I. Barmpadimos, C. Hueglin, J. Keller, S. Henne, and A. Prévôt, "Influence of meteorology on PM10 trends and variability in Switzerland from 1991 to 2008," *Atmospheric Chemistry and Physics,* vol. 11, no. 4, pp. 1813-1835, 2011.
[48] N. Masey, J. Gillespie, M. Heal, S. Hamilton, and I. Beverland, "Influence of wind-speed on short-duration NO2 measurements using Palmes and Ogawa passive diffusion samplers," *Atmospheric Environment,* vol. 160, no. 1, pp. 70-76, 2017.
[49] M. Sjöholm, "Dispersion Pattern Of Nitrogen Dioxide Within A Dense Urban Structure During Different Meteorology," Master Thesis, Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden, 2017.
[50] M. Aljanabi, M. Shkoukani, and M. Hijjawi, "Ground-level ozone prediction using machine learning techniques: A case study in Amman, Jordan," *International Journal of Automation Computing,* vol. 17, pp. 667-677, 2020.