Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan

Article *in* International Journal of Automation and Computing · May 2020

DOI: 10.1007/s11633-020-1233-4

3 authors:

Maryam Aljanabi
Applied Science Private University
5 PUBLICATIONS **76** CITATIONS
SEE PROFILE

Mohammad Shkoukani
Applied Science Private University
24 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Mohammad Hijjawi
Applied Science Private University
20 PUBLICATIONS **188** CITATIONS
SEE PROFILE

International Journal of Automation and Computing

DOI: 10.1007/s11633-020-1233-4

Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan

Maryam Aljanabi Mohammad Shkoukani Mohammad Hijjawi

Faculty of Information Technology, Applied Science Private University, Amman 11931, Jordan

Abstract: Air pollution is one of the most serious hazards to humans′ health nowadays, it is an invisible killer that takes many human lives every year. There are many pollutants existing in the atmosphere today, ozone being one of the most threatening pollutants. It can cause serious health damage such as wheezing, asthma, inflammation, and early mortality rates. Although air pollution could be forecasted using chemical and physical models, machine learning techniques showed promising results in this area, especially artificial neural networks. Despite its importance, there has not been any research on predicting ground-level ozone in Jordan. In this paper, we build a model for predicting ozone concentration for the next day in Amman, Jordan using a mixture of meteorological and seasonal variables of the previous day. We compare a multi-layer perceptron neural network (MLP), support vector regression (SVR), decision tree regression (DTR), and extreme gradient boosting (XGBoost) algorithms. We also explore the effect of applying various smoothing filters on the time-series data such as moving average, Holt-Winters smoothing and Savitzky-Golay filters. We find that MLP outperformed the other algorithms and that using Savitzky-Golay improved the results by 50% for coefficient of determination (R^2) and 80% for root mean square error (RMSE) and mean absolute error (MAE). Another point we focus on is the variables required to predict ozone concentration. In order to reduce the time required for prediction, we perform feature selection which greatly reduces the time by 91% as well as shrinking the number of features required for prediction to the previous day values of ozone, humidity, and temperature. The final model scored 98.653% for R^2, 1.016 ppb for RMSE and 0.800 ppb for MAE.

Keywords: Ozone prediction, machine learning, neural networks, supervised learning, regression.

1 Introduction

Air pollution is one of the major hazards to human health and the ecosystem nowadays. With the growing development of the economy, the rising population, the increase in industries and the growing need for transportation, this all leads to increased environmental pollution which includes air pollution. Air pollution is mainly caused by the emissions of factories, electrical facilities, vehicles that burn fuel, meteorological factors, etc.[1] Ground-level ozone is a major pollutant that is hazardous to humans′ health, unlike the stratospheric ozone layer that protects the earth. It is formed as a reaction between pollutants resulting from industrial emissions, vehicles, and electrical facilities. Health problems associated with ozone exposure include wheezing, coughing, asthma, chest pain, decreased capacity for exercise, inflammation, increased mortality rate and more[2]. It can have a severe impact not only on humans but also on vegetation and crops. It caused a €6.7 billion crop loss in the EU in 2007[3]. Ozone is considered one of the greenhouse gases that causes a reduction of carbon intake by plants which contributes to increased global warming[3]. Due to the dangers of air pollution, multiple air quality indices exist in different countries and they are used to determine if the pollutants′ concentrations are within the healthy range[4].

With the decreased cost of pollutants′ monitoring sensors, many projects are being carried out in different countries to monitor pollutants, e.g., stations measuring pollutants which are connected to the internet of things.

Some of these projects have been collecting and storing data for several years which can lead to more knowledge about the problem of air pollution[5, 6]. Since the issue of air quality is of high significance, there have been multiple attempts to forecast air quality in different methods. Air quality forecasting systems are tools that can help describe the air quality problem and understand the relationship between pollutants, meteorological factors, emissions, and other atmospheric variables. They can help make future forecasts about air quality[1]. Types of forecasting systems include deterministic models that use mathematical equations to describe the atmospheric processes causing pollution. They are based on the physical and chemical nature of pollutants. The problem with these chemical and physical models is that they do not capture the behavior of pollutants very well and they tend to linearize a non-linear relationship between data in the natural world as well as having difficulties in processing large amounts of data[7]. Due to the problems of

Research Article

Manuscript received December 20, 2019; accepted April 7, 2020

Recommended by Associate Editor Paul Stewart

the afore-mentioned models and the increase in available data, new methods were introduced to discover patterns in data and make better predictions for air pollutants. Machine learning, which is a subfield of artificial intelligence could be used in this case due to its ability to discover complex relationships between data and to analyze large datasets[5]. One of the most well-known machine learning algorithms is artificial neural networks (ANN) which is widely used by many authors due to its ability to discover non-linear relationships between variables[8]. Another algorithm that is used in this topic is the support vector machine (SVM) which is also called support vector regression (SVR) when used for regression purposes[9]. This algorithm is a good generalization algorithm that generalizes well to new data[10]. The third algorithm that we used is the decision tree (DT) which is a well-known machine learning algorithm with a graphical upside-down tree structure. When a decision tree is used for regression, it is called a decision tree regression (DTR) or regression tree[11]. Finally, extreme gradient boosting (XGBoost) which is a boosted tree with the gradient boosting method is being used for prediction purposes due to its promising results and speed[12].

Despite the fact that many studies were conducted in various parts of the world to forecast ozone concentration using machine learning, the problem of air pollution prediction is not given significant importance in Jordan and thus no papers were done on this topic. Another point is papers in the context of air quality forecasting rarely focused on time-series smoothing filters and the time required for prediction. In this paper, we tackled the problem of ozone prediction in Jordan using machine learning techniques. We focused on three main topics in this paper: Firstly, we conducted a comparison between four machine learning algorithms which are multi-layer perceptron (MLP), SVR, DTR, and XGBoost to find the algorithm with the highest performance. Secondly, we explored the importance of adding a smoothing filter to the noisy time-series dataset. Previous research in [13] explained how using a denoising filter improved the results in the field of air pollution prediction. In this paper, we performed a comparison between three smoothing filters and compared their results with the original unfiltered data. Thirdly, we tried to decrease the time required for prediction by finding the most important features for predicting ozone concentration, since many variables in the dataset may not be relevant to the prediction process. Although ozone is highly affected by many complex atmospheric and meteorological variables, conducting feature selection and narrowing down the number of features proved efficient and greatly reduced the time and improved the results.

This paper is structured as follows. The related work section contains a brief explanation of ANN, SVR, DTR, and XGBoost alongside previous research done in the field of air quality forecasting using these algorithms. The materials and methods section displays information about the dataset as well as describing some concepts about the smoothing filters and the performance evaluation metrics used in this research. The experimental results and discussion section illustrates each step of the experiments in detail such as the parameters configurations of the used algorithms, the results obtained, and a discussion of the results. Finally, the conclusion and related work section summarizes this research and gives ideas for future research work in this field.

2 Related work

Multiple machine learning algorithms were used for building ozone concentration forecasting systems all around the world. One of the most widely used algorithms, however, is ANN which proved efficient not only in the topic of air quality prediction but in many other topics as well. ANN did not only show superior performance in the field of air quality prediction, but also in other natural world forecasting systems such as wind speed prediction[14] and rainfall prediction[15]. The ANN is an algorithm that tries to mimic how the neurons work in the human brain and learn the way they process information. The basic computing unit of the ANN is the neuron (also called the perceptron), which is represented by a circle, that receives multiple inputs and produces a single output. ANN has many types, such as multi-layer perceptron (MLP), radial basis function (RBF), recurrent neural networks (RNN), convolutional neural networks (CNN), etc. The ANN generally consists of an input layer, hidden layers, and an output layer. The neurons in the input layer represent the input features (X) of the dataset. The hidden layer is where the complexity lies and where the learning happens. The output layer represents the output of the network[16, 17]. It contains a single neuron in the case of ozone concentration prediction, which is the numerical value representing ozone concentration.

Abdul-Wahab and Al-Alawi developed an ozone prediction model in Kuwait using ANN in [5]. The input variables to the system were a combination of meteorological variables as well as other pollutants that existed in the dataset. The research explored the relationship between ozone and other variables in the dataset and proved that meteorological variables contributed to ozone concentration by 33.15% to 40.64%. Prybutok et al.[18] proved that ANN outperformed classical statistical models for predicting ozone. The inputs to the developed model contained pollutants like carbon monoxide, nitrogen dioxide, nitric oxide, etc. as well as meteorological variables. The research showed that ANN outperformed the regression model, and the autoregressive integrated moving average (ARIMA) model and scored the lowest values for mean absolute deviation (MAD) and root mean square error (RMSE).

Faris et al.[19] investigated predicting surface ozone levels. This model's inputs were meteorological variables with one pollutant which is nitrogen dioxide. The authors compared MLP and RBF types of ANN and showed that MLP had the lowest error rates. Another type of ANN called cyclic reservoir with jumps (CRJ) was used by Sheta et al.[20] The research focused on predicting ozone concentration in eastern Croatia, namely in Osijek city and Kopački. The system inputs contained meteorological variables, PM10 values and ozone concentration of the previous day. The model showed promising results for CRJ. Another study focused on predicting ozone through seasonal relation by Kumar et al.[21] The inputs to the model were three variables only which were nitrogen dioxide, temperature and humidity. The authors compared three types of ANN which are MLP, RBF as well as and generalized regression neural network (GRNN). The model that outperformed the others was MLP for all seasons and scored the lowest RMSE and mean absolute error (MAE). The lowest error values were found during the winter season.

MLP also showed promising results in research conducted by Pauli et al.[22] in Corsica in the Genova gulf. The focus of the research was experimenting with different sets of input variables to the system and determining the best combination of variables. The variables were a combination of meteorological variables, other pollutants, seasonal variables, and the same pollutant with different time lag values. Extreme machine learning (ELM) is another type of ANN that was used in [1] to predict ozone, $PM_{2.5}$, and nitrogen dioxide in six Canadian cities. The researcher used a mixture of hourly meteorological predictors as well as persistence, chemical and physical predictors to predict the above-mentioned pollutants. The result showed that using a method named online-sequential extreme learning machine (OS-ELM) showed an improvement in most cities. In the case of big data containing tens or hundreds of thousands of records, deep learning neural networks showed good results. For example, Li et al.[23] developed spatio-temporal deep learning (STDL) to forecast $PM_{2.5}$ in Beijing, China. The dataset contained 20 196 records collected from 12 stations. The authors comparing STDL with SVR, auto regression moving average (ARMA), and spatiotemporal artificial neural network (STANN), and STDL showed superior results to them.

SVR was also used but less frequently than ANN in the topic of air quality prediction. This algorithm tries to maximize the margin between the boundary points (also called the support vectors) to minimize the errors. It uses various parameters such as a kernel function to achieve its goal and there are multiple kernel functions to choose from depending on the problem at hand[24]. Yet choosing the optimal parameters of SVR can be hard to determine[25]. Wang et al.[26] developed a model to predict respirable suspended particulates (RSP) using SVR in Hong Kong, China. The comparison was between feeding the data sequentially into the SVR model which was called online SVR, and feeding it in batch mode in the normal SVR model. The research concluded that the online SVR model was superior. In another work by Liu et al.[27], SVR was also used to forecast RSP concentration and it was compared to RBF. The experiments proved that SVR outperformed RBF in the research.

DTR is a rule-based machine learning technique that is widely used in prediction models. It describes the relationships between the variables in a tree structure[28]. XGBoost which is an enhanced tree-based algorithm also showed promising results and speed in various works. It is a relatively new algorithm and less frequently used than SVR and ANN. XGBoost was used to forecast $PM_{2.5}$ in Tianjin, China by Pan[29]. The hourly data included pollution features like ozone, nitrogen dioxide, sulfur dioxide, carbon monoxide, and PM_{10}. Compared with various machine learning models like SVR, decision tree, random forest, and multiple linear regression, XGBoost showed better performance and reduced error values than the previously mentioned algorithms. Joharestani et al.[30] also predicted daily $PM_{2.5}$ in Tehran, Iran using XGBoost. The research showed that XGBoost outperformed random forest and deep learning methods and demonstrated the lowest errors. The research also explored feature importance and found that a lag variable of one day for $PM_{2.5}$ was the most valuable in the prediction process.

From the above-mentioned research, we can see that researchers used various ANN, SVR, XGBoost structures as well as different dataset averaging, either hourly or daily. Another point is the number and the nature of the input variables to the system. Some researchers used only meteorological variables while others used a mixture of the ozone levels of the previous day or days with the meteorological variables. In some research, the authors went ahead and tried adding other pollutants to the input model as well to forecast ozone concentration[31].

3 Methodology

3.1 Dataset and area description

Amman is the capital of Jordan. It's a fast-growing Arab city with increased urbanization and economic growth and is currently one of the most important cities in the Arab world. Amman is located in the north-west of Jordan as seen in Fig. 1. The topography of the city consists of a lot of hills and valleys. Amman's climate consists of four seasons, although the autumn and spring are relatively short. The summer is a rainless season with low humidity, while winter is a cold season with temperature often going below zero and heavy rainfall in January and February. However, these patterns may vary due to climate change issues that are affecting the whole globe[32].

The dataset of this study was obtained from King Hussein Public Park station which is located in a residen-

Fig. 1 Map of Jordan[33]

tial area in Amman. The dataset covers the period from May 1st, 2014, to June 4th, 2019, and was obtained from the Jordanian Ministry of Environment. The dataset contained daily average readings of ozone (ppb) as well as meteorological variables like relative humidity (%), ambient temperature (°C), wind speed (km/h), and wind direction (°)[34]. The statistical description of the dataset variables is presented in Table 1. Furthermore, a scatterplot demonstrating the effect of meteorological variables on ozone is presented in Fig. 2. We can see how ozone levels increase as the values of temperature, wind speed, and wind direction increase. On the other hand, it has an inverse relationship with humidity, meaning that ozone concentrations decrease as humidity increase.

Table 1 Statistical description of the dataset features

Feature	Mean	Standard deviation	Min	Max
Ozone	39.943	11.505	1.1700	70.800
Temperature	15.637	6.847	0.280	31
Humidity	63.309	20.301	21.471	100
Wind direction	223.219	54.685	48.125	324
Wind speed	11.0465	7.116	2.070	41.800

The dataset was preprocessed to prepare it for the machine learning algorithm and maximize the prediction performance. The first step in the data processing for time-series is feature engineering, which involves adding meaningful date variables to the dataset to assist in the prediction process. In our case, we added a special day feature which shows if a day is a weekend/holiday or not. It takes one of two values, 0 or 1. The second feature is the day of the year feature, this feature takes values from 1 to 365 and it can reflect various seasonal characteristics in the prediction process.

The second step is treating missing values. Our dataset contained a total of 131 missing values including 4 for ozone, 10 for humidity, 23 for temperature, 47 for wind speed, and 47 for wind direction. One way for treating missing values in time-series data is using interpolation, which is a mathematical method for filling the missing values using a function whether it is linear, polynomial, etc. The problem with interpolation is when there are missing values at the beginning of the dataset, it cannot fill them properly. So we removed the first month because it contained a lot of missing values, which left us with the interval from June 4th, 2014, to June 4th, 2019, which is exactly 5 years with 1826 records. The dataset was also normalized using the MinMax scaler which transforms the dataset values to lie between 0 and 1. In the experiments when we used the smoothing filters, normalization was done after applying the filter. We used 60% of the data for training, which is 3 years, while 40% was used for testing which corresponds to 2 years.

3.2 Preprocessing smoothing filters

Time-series data tends to be noisy and contains a lot of fluctuations which could negatively affect the performance of machine learning algorithms. This makes the noise removal stage, called denoising, the most vital stage in the preprocessing steps because it can transform the noisy time-series data into smooth data without losing any information in the process[35]. Since the effect of smoothing filters is not investigated in the case of ozone prediction, in this research we conducted a comparison between three filters, Holt-Winters, moving average, and Savitsky-Golay.

Holt-Winters is an exponential smoothing method that recognizes seasonal patterns in the data. It is usually used to forecast time-series data since most time-series data exhibit seasonal patterns[36]. Yet in this research, we are using it as a smoothing technique since it can eliminate noise. The moving average smoothing filter works by averaging a certain number of points to produce a new one. The number of points used for averaging is chosen by the researcher. This filter often produces good results for problems like noise elimination since if we have multiple noisy samples, then not all of them are relevant and thus they can be averaged to reduce their random noise. The larger the number used for averaging, the smoother the data becomes, which is not always correct[37]. The number must be chosen so that the data still preserves its shape yet the noise is eliminated. In our research, we used an averaging number of 7. The Savitsky-Golay filter is a low-pass filter that provides a method for smoothing time-series data using local least-squares polynomial approximation smoothing technique[38]. Savitsky-Golay was used to preprocess the data used to predict $PM_{2.5}$ in [13] and improved the results drastically. It was also compared to wavelet analysis and proved superior to it. Furthermore, the successful use of Savitsky-Golay was also demonstrated in [39–42] to

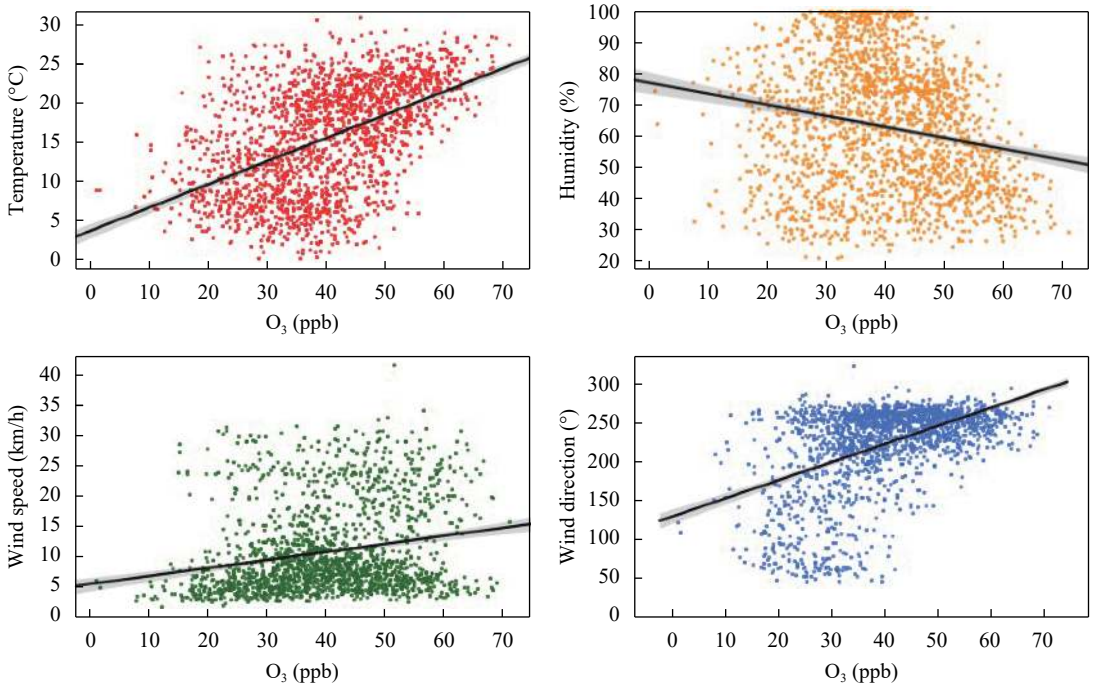

Fig. 2 A scatterplot of ozone versus the meteorological variables in the dataset

smooth noisy data and solve the problem when the amount of noise in the dataset becomes a hindrance to the prediction process.

3.3 Performance evaluation metrics

In this research, we are going to predict the numeric concentration of ozone, which is a regression model in machine learning. The model evaluation metrics that we are going to use are as follows:

Coefficient of determination

The coefficient of determination explains the variation between two variables in the case of a regression model. It shows the relationship between the actual value and the predicted value. It ranges between 0 and 1, where 1 is the perfect correlation between the actual and the predicted output (highly in agreement) and 0 means they have no correlation. Equation (1) can be used to calculate it:

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

where N is the number of samples, A_i is the actual value, P_i is the predicted value, $\bar{A}$ is the average of the actual values, P is the average of the predicted values, σ_P is the standard deviation of the predicted values and σ_A is the standard deviation of the actual values[13].

Root mean square error

The root mean square error shows the root of the mean squared error between the actual value and the predicted value of the regression model. Equation (2) shows how it can be calculated:

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

where N is the number of samples, P_i is the predicted value, A_i is the actual value[43].

Mean absolute error

The mean absolute error illustrates the average of the absolute errors in the regression model. It can be explained in (3):

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |P_i - A_i| . \tag{3}$$

Its value is usually less than the RMSE since it takes the absolute value without squaring it and it is less sensitive to extreme error values than RMSE[43].

4 Experimental results and discussion

All the experiments were carried out using python 3. The machine used to run the experiments was a Windows 8.1 64-bit HP laptop with a Core-i5 processor, 2.2 GHz, and 4 GB RAM.

4.1 Models building and parameter optimization

This step involves building the MLP, SVR, DTR, and

XGBoost models. The optimal parameters for all the models are shown in Table 2. They were found through trying different combinations of parameters. The inputs to the models at this stage contain the normalized unfiltered previous day's values of ozone, temperature, humidity, wind speed, wind direction, as well as the special day and the day of the year. The results shown in Table 3 shows that MLP was superior to the other models with an R^2 of 65.505%, followed by XGBoost and SVR, with DTR having the lowest performance of 60.794% for R^2. Based on these results, the MLP model results will be used as a benchmark for the next step which is comparing the different smoothing filters.

Table 2 Optimal parameter configurations for MLP, SVR, DTR, and XGBoost models

Model	Parameters
MLP	One hidden layer with 150 neurons, solver: "adam", activation function: "tanh"
SVR	Kernel function : "rbf", gamma: 0.01, C : 10, epsilon: 0.001
DTR	Max depth: 3, min_samples_split=0.1, min_samples_leaf=0.1
XGBoost	Booster: "gbtree", max-depth: 2, min-child-weight: 5, learning-rate: 0.06

Table 3 Comparison between MLP, SVR, DTR, and XGBoost to predict ozone concentration

Model	R^2 (%)	RMSE (ppb)	MAE (ppb)
MLP	65.505	6.077	4.717
SVR	62.637	6.325	4.966
DTR	60.794	6.479	5.095
XGB	63.493	6.252	4.937

In the next step, we applied the three filters that we mentioned earlier, which are Holt-Winters smoothing, moving average filter with 7 days averaging, and Savitzky-Golay filter. As seen in Table 4, all filters improved the result yet the filter that outperformed the others was Savitzky-Golay. It yielded an R^2 of 98.231% which is a great improvement of about 50% in the performance of R^2 as compared to the benchmark result of 65.505% from before applying any filter.

The RMSE and MAE are improved by about 80%. We experimented with different window length and polynomial for the Savitzky-Golay filter till we arrived at the best combination for our data which happened to be 25 for the window length and 4 for the polynomial. This combination smoothed the data but at the same time no information was lost and the data still kept its shape. All the MLP models have a single hidden layer of 140, 140, and 270 neurons for the Holt-Winters, moving average, and Savitzky-Golay filters respectively which were obtained through a process of trial and error. All the models used "tanh" as the activation function and "lbfgs" as the solver except for the Savitzky-Golay filter that uses the "relu" activation function. Note that when we used the smoothing filters, we normalized the data after using the filter.

Table 4 MLP results for the dataset before and after the filters

Model	R^2 (%)	RMSE (ppb)	MAE (ppb)
No filter (original MLP)	65.505	6.077	4.717
Holt-Winters	89.792	2.938	2.311
Moving average	97.804	1.288	1.012
Savitzky-Golay	98.231	1.165	0.917

We can see how applying the Savitzky-Golay filter smoothed the data, decreased the noise, and even removed the outliers in the dataset. Figs. 3–7 show the effect of the filter with the light line being the original data and the dark line being the filtered data.

4.2 Feature selection results

In this step, we focused on the time of prediction. The prediction error is always the determining factor in selecting the number of features, however, when the errors are very close to each other and barely different, we can rely on time to find the best feature combinations that could achieve the topmost performance. When the dataset size grows or when using hourly values, the time is essential as well as the number of features. A lower number of features means lower computational resources, decreased time and cost. Feature selection is the process of finding the most relevant features required for prediction and thus reducing the time and the computational resources needed. We used the forward wrapper feature selection method with the MLP model on the data after Savitzky-Golay smoothing step. The wrapper method tries to find the best subset of features that yields the best performance. Since we have 7 features, and since the forward wrapper required a number of features as one of its parameters, we tried 6, 5, 4 and 3 features and each time we built a new MLP model to test if this combination of features is the optimal one or not. The result is shown in Table 5.

Although the different combinations of features exhibited minor enhancements in terms of performance metrics, yet the 3 features achieved the best results even if the improvement was slight, and it also reduced the time by about 91% from 434 ms to 37 ms which is very beneficial in the case of big datasets. Note that all features in the features column in Table 5 are the values of the previous day.

We can see that ozone, humidity, and temperature of the previous day are the variables that affect ozone levels of the next day the most. This makes sense since ozone increases in the summer months when the temperature is

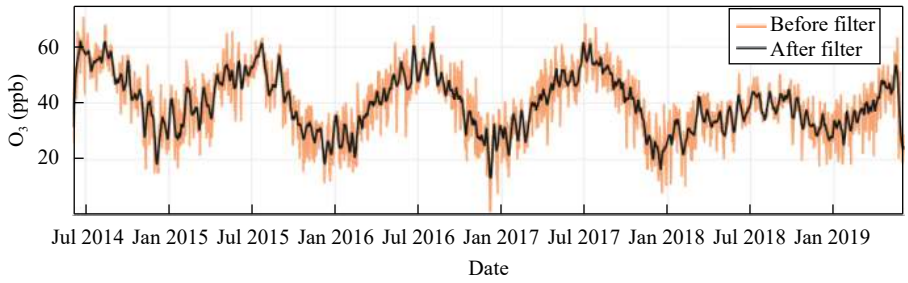

Fig. 3 O_3 data before and after applying filter

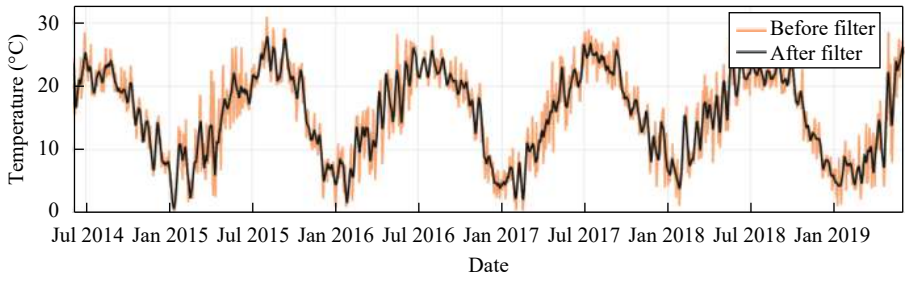

Fig. 4 Temperature data before and after applying filter

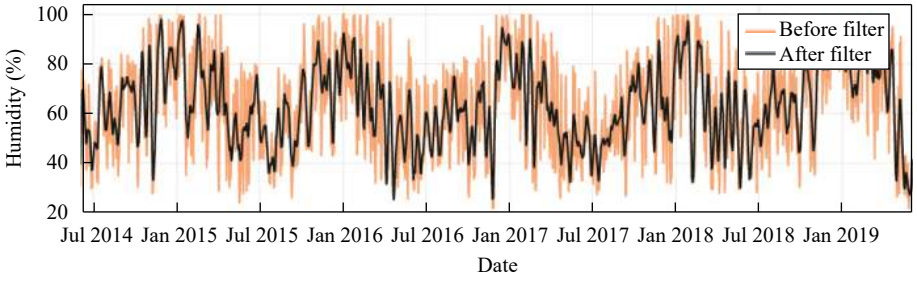

Fig. 5 Humidity data before and after applying filter

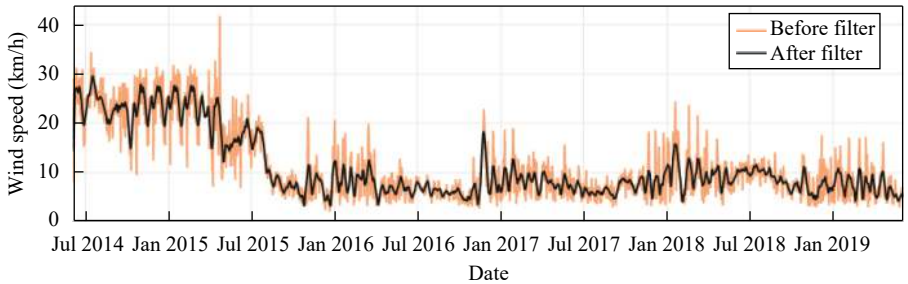

Fig. 6 Wind speed data before and after applying filter

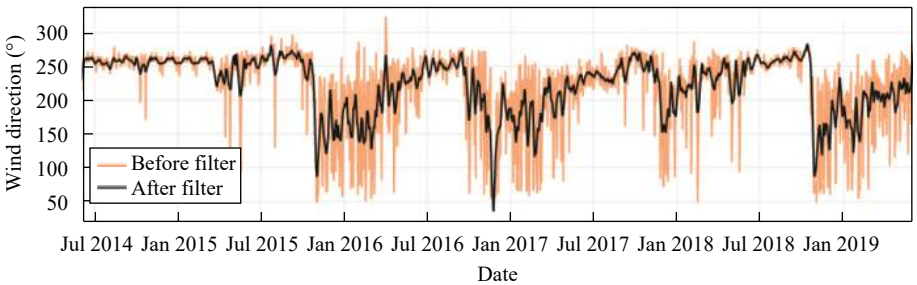

Fig. 7 Wind direction data before and after applying filter

high and when humidity is low. We also notice that wind speed and wind direction are irrelevant and removing them improved the prediction result. We can deduce that three variables are enough to predict ozone in our data-

Table 5 Feature selection results

Number	Features	R^2 (%)	RMSE (ppb)	MAE (ppb)	CPU Time (ms)	MLP Parameters
All	Ozone, temperature, humidity, wind speed, wind direction, special day, and day of the year	98.231	1.165	0.917	434	One hidden layer, 270 neurons,activation function: "relu"
6	Ozone, temperature, humidity, wind direction, special day, and day of the year	98.286	1.146	0.905	239	One hidden layer, 140 neurons,activation function: "relu"
5	Ozone, temperature, humidity, special day, and day of the year	98.537	1.059	0.829	84	One hidden layer, 100 neurons,activation function: "tanh"
4	Ozone, temperature, humidity, and day of the year	98.639	1.021	0.802	62	One hidden layer, 60 neurons,activation function: "tanh"
3	Ozone, temperature, and humidity	98.653	1.016	0.800	37	One hidden layer, 15 neurons,activation function: "relu"

set and thus excluding the rest of the features not only improved the results but also reduced the time. Another point to make is using three variables can reduce the cost required for prediction, as the final model only requires ozone, humidity and temperature readings to forecast ozone levels of the next day.

The final model obtained used 3 neurons in the input layer, 15 neurons in the hidden layer and one neuron in the output layer to predict ozone concentration of the next day. Fig. 8 illustrates the MLP configuration of the built model. This combination was found based on trying multiple neural network configurations. Upon trying, we found that one hidden layer always showed better and faster results than multiple layers which did not enhance the results. It was also noted that as the number of features decreased in the feature selection step, the number of neurons in the hidden layer also decreased, which contributed to a major enhancement in time. Note that in Fig. 8, all the features in the input layer refer to features' concentrations of the previous day.

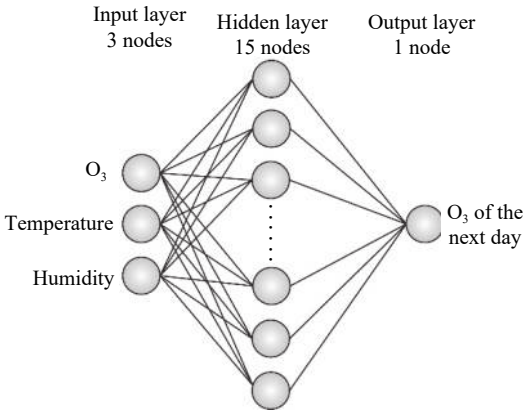

Fig. 8 MLP model architecture

The results shown in Table 5 are illustrated in Figs. 9–12 to show the change in the performance evaluation metrics as well as to demonstrate the drop in time. We can clearly observe how the 3 features outperform the other feature combinations.

Finally, Fig. 13 shows the actual and predicted ozone graph. We can see that the predicted result is almost identical to the actual with a small error rate.

5 Conclusions and future work

In this research, we assessed the ability of machine learning techniques to predict ozone levels for the next day in Amman, Jordan, specifically, in King Hussein Public Parks and the surrounding area. We compared MLP, SVR, DTR, and XGboost and found that MLP outperformed the other algorithms. We also compared various smoothing filters for the time-series data and discovered that the Savitsky-Golay filter enhanced the results by 50% for R^2 and 80% for both RMSE and MAE. The final contribution of this research is performing an intensive feature selection to reduce the number of features and thus decrease the time it takes to make the prediction since time is an important factor in the case of large datasets. Using the forward wrapper, we found that the previous day values of ozone, temperature, and humidity are the most influential features in our dataset for forecasting ozone concentration of the next day. The time is improved from before and after using the feature selection by about 91%. The final developed model scored R^2 of 98.653%, RMSE of 1.016 ppb and MAE of 0.800 ppb which is a very promising result.

For future work, we suggest using hourly data if possible to see if MLP would still outperform the others, or even experiment with different deep learning neural networks since they are best suited for large amounts of data. Long short-term memory (LSTM) neural network would be suitable in this case since it is a deep learning model that deals with predicting time-series datasets. Another point is to use different meteorological or pollution variables that may improve the prediction or prove of high importance to ozone prediction and thus lead to different results in the feature selection phase. It would also be a good idea to try this model on a dataset from another country to see the difference in results.

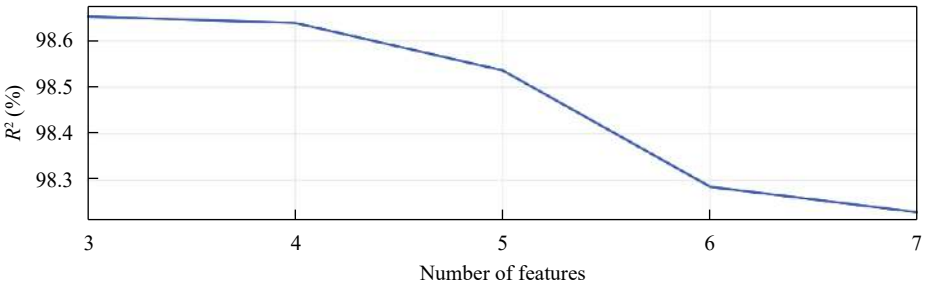

Fig. 9 R^2 versus the number of features

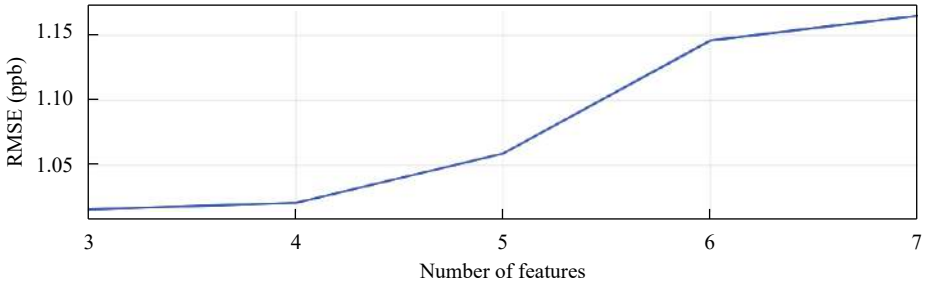

Fig. 10 RMSE versus the number of features

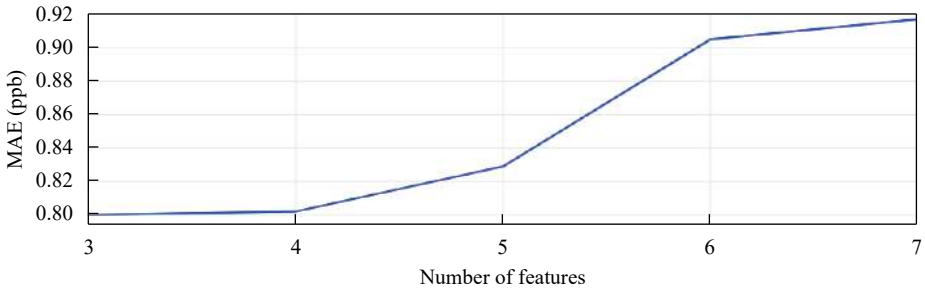

Fig. 11 MAE versus the number of features

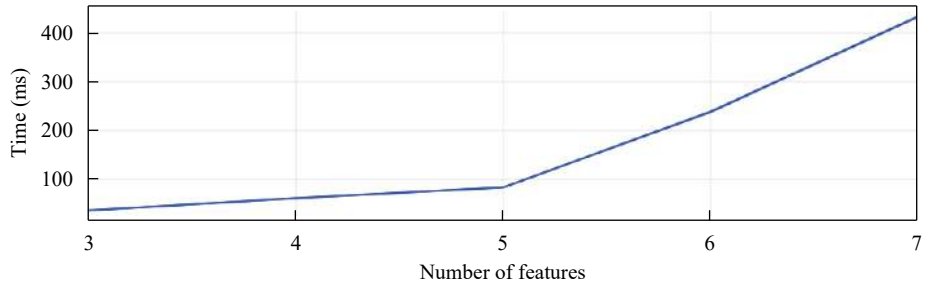

Fig. 12 Time versus the number of features

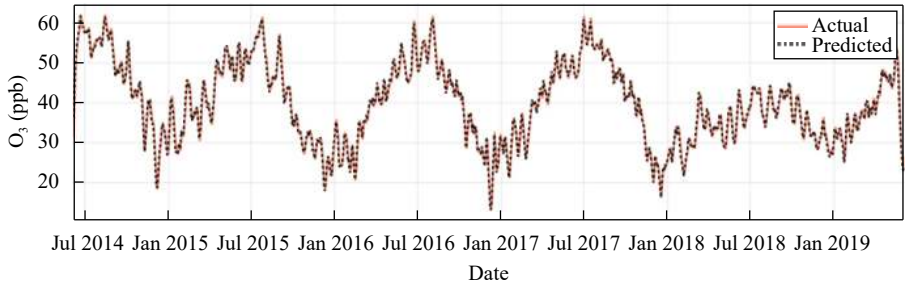

Fig. 13 Actual and predicted ozone levels

Acknowledgements

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the financial support granted to this research. The authors would also like to thank the Jordanian Ministry of Environment for giving them access to perform scientific research on the King Al-Hussein Public Parks air pollution and meteorological data collected by the ministry.

References

[1] H. P. Peng. Air Quality Prediction by Machine Learning Methods, Master dissertation, The University of British Columbia, Canada, 2015.

[2] United States Environmental Protection Agency. Environments and contaminants: Criteria air pollutants. *America's Children and the Environment*, 3rd ed., United States Environmental Protection Agency, Ed., Washington DC, USA: United States Environmental Protection Agency, 2015.

[3] DEFRA. *Air Pollution: Action in a Changing Climate*, London, UK: Department for Environment, Food and Rural Affairs, 2010.

[4] A. Plaia, M. Ruggieri. Air quality indices: A review. *Reviews in Environmental Science and Bio/Technology*, vol. 10, no. 2, pp. 165–179, 2011. DOI: 10.1007/s11157-010-9227-2.

[5] C. Bellinger, M. Shazan, M. Jabbar, O. Zaïane, A. Osornio-Vargas. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, vol. 17, no. 1, Article number 907, 2017. DOI: 10.1186/s12889-017-4914-3.

[6] T. M. Chiwewe, J. Ditsela. Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations. In *Proceedings of the 14th IEEE International Conference on Industrial Informatics*, IEEE, Poitiers, France, pp. 58-63, 2016. DOI: 10.1109/INDIN.2016.7819134.

[7] S. A. Abdul-Wahab, S. M. Al-Alawi. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environmental Modelling & Software*, vol. 17, no. 3, pp. 219–228, 2002. DOI: 10.1016/S1364-8152(01)00077-9.

[8] W. Z. Lu, D. Wang. Learning machines: Rationale and application in ground-level ozone prediction. *Applied Soft Computing*, vol. 24, pp. 135–141, 2014. DOI: 10.1016/j.asoc.2014.07.008.

[9] A. S. Sánchez, P. J. G. Nieto, P. R. Fernández, J. J. del Coz Díaz, F. J. Iglesias-Rodríguez. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Mathematical and Computer Modelling*, vol. 54, no. 5-6, pp. 1453–1466, 2011. DOI: 10.1016/j.mcm.2011.04.017.

[10] A. J. Smola, B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004. DOI: 10.1023/B:STCO.0000035301.49549.88.

[11] G. G. Moisen. Classification and regression trees. *Encyclopedia of Ecology*, S. E. Jørgensen, B. D. Fath, Eds., Oxford, UK: Elsevier, 2008.

[12] B. X. Zhai, J. G. Chen. Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China. *Science of the Total Environment*, vol. 635, pp. 644–658, 2018. DOI: 10.1016/j.scitotenv.2018.04.040.

[13] M. R. Delavar, A. Gholami, G. R. Shiran, Y. Rashidi, G. R. Nakhaeizadeh, K. Fedra, S. Hatefi 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*, vol. 8, no. 2, Article number 99, 2019. DOI: 10.3390/ijgi8020099.

[14] S. P. Mishra, P. K. Dash. Short term wind speed prediction using multiple kernel pseudo inverse neural network. *International Journal of Automation and Computing*, vol. 15, no. 1, pp. 66–83, 2018. DOI: 10.1007/s11633-017-1086-7.

[15] S. R. Devi, P. Arulmozhivarman, C. Venkatesh, P. Agarwal. Performance comparison of artificial neural network models for daily rainfall prediction. *International Journal of Automation and Computing*, vol. 13, no. 5, pp. 417–427, 2016. DOI: 10.1007/s11633-016-0986-2.

[16] S. Haykin. *Neural Networks: A Comprehensive Foundation*, 3rd ed., Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2007.

[17] V. K. Ha, J. C. Ren, X. Y. Xu, S. Zhao, G. Xie, V. Masero, A. Hussain. Deep learning based single image super-resolution: A survey. *International Journal of Automation and Computing*, vol. 16, no. 4, pp. 413–426, 2019. DOI: 10.1007/s11633-019-1183-x.

[18] V. R. Prybutok, J. Yi, D. Mitchell. Comparison of neural network models with ARIMA and regression models for prediction of Houston′s daily maximum ozone concentrations. *European Journal of Operational Research*, vol. 122, no. 1, pp. 31–40, 2000. DOI: 10.1016/S0377-2217(99)00069-7.

[19] H. Faris, M. Alkasassbeh, A. Rodan. Artificial neural networks for surface ozone prediction: Models and analysis. *Polish Journal of Environmental Studies*, vol. 23, no. 2, pp. 341–348, 2014.

[20] A. Sheta, H. Faris, A. Rodan, E. Kovač-Andrić, A. M. 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. DOI: 10.1007/s11869-018-0561-9.

[21] N. Kumar, A. Middey, P. S. Rao. Prediction and examination of seasonal variation of ozone with meteorological parameter through artificial neural network at NEERI, Nagpur, India. *Urban Climate*, vol. 20, pp. 148–167, 2017. DOI: 10.1016/j.uclim.2017.04.003.

[22] C. Paoli, G. Notton, M. L. Nivet, M. Padovani, J. L. Savelli. A neural network model forecasting for prediction of hourly ozone concentration in corsica. In *Proceedings of the 10th International Conference on Environment and Electrical Engineering*, IEEE, Rome, Italy, 2011. DOI: 10.1109/EEEIC.2011.5874661.

[23] X. Li, L. Peng, Y. Hu, J. Shao, T. H. Chi. Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, vol. 23, no. 22, pp. 22408–22417, 2016. DOI: 10.1007/s11356-016-7812-9.

[24] T. X. Zhang, J. Y. Su, C. J. Liu, W. H. Chen. Potential bands of sentinel-2A satellite for classification problems in precision agriculture. *International Journal of Automation and Computing*, vol. 16, no. 1, pp. 16–26, 2019. DOI: 10.1007/s11633-018-1143-x.

[25] Z. D. Tian, X. W. Gao, K. Li. A hybrid time-delay predic-

tion method for networked control system. *International Journal of Automation and Computing*, vol. 11, no. 1, pp. 19–24, 2014. DOI: 10.1007/s11633-014-0761-1.

[26] W. J. Wang, C. Q. Men, W. Z. Lu. Online prediction model based on support vector machine. *Neurocomputing*, vol. 71, no. 4–6, pp. 550–558, 2008. DOI: 10.1016/j.neucom.2007.07.020.

[27] B. C. Liu, A. Binaykia, P. C. Chang, M. K. Tiwari, C. 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, Article number e0179763, 2017. DOI: 10.1371/journal.pone.0179763.

[28] M. S. Tehrany, B. Pradhan, M. N. Jebur. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *Journal of Hydrology*, vol. 504, pp. 69–79, 2013. DOI: 10.1016/j.jhydrol.2013.09.034.

[29] B. Y. Pan. Application of XGBoost algorithm in hourly $PM_{2.5}$ concentration prediction. In *IOP Conference Series: Earth and Environmental Science*, vol. 113, Article number. 012127, 2018. DOI: 10.1088/1755-1315/113/1/012127.

[30] M. Z. Joharestani, C. X. Cao, X. L. Ni, B. Bashir, S. Talebiesfandarani. $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, vol. 10, no. 7, Article number 373, 2019. DOI: 10.3390/atmos10070373.

[31] Y. Rybarczyk, R. Zalakeviciute. Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, vol. 8, no. 12, Article number 2570, 2018. DOI: 10.3390/app8122570.

[32] R. B. Potter, K. Darmame, N. Barham, S. Nortcliff. An Introduction to the Urban Geography of Amman, Jordan. Geographical Paper No. 182, The University of Reading, UK, 2007.

[33] Stamen and OpenStreetMap. Stamen Maps, [Online], Available: http://maps.stamen.com/toner/#6/31.588/35.552, February 20, 2020.

[34] Jordanian Ministry of Environment. Daily Pollution Concentrations in King Al-Hussein Public Parks Station Dataset, [Online], Available: http://moenv.gov.jo/EN/Pages/mainpage.aspx, 2019.

[35] R. M. Alrumaih, M. A. 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. DOI: 10.1016/S1018-3639(18)30755-4.

[36] A. M. De Livera, R. J. Hyndman, R. D. 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. DOI: 10.1198/jasa.2011.tm09771.

[37] S. W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd ed., San Diego, USA: California Technical Publishing, 1999.

[38] R. W. Schafer. What is a savitzky-golay filter? *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111–117, 2011. DOI: 10.1109/MSP.2011.941097.

[39] S. B. Ashrafi, M. Anemangely, M. Sabah, M. J. Ameri. Application of hybrid artificial neural networks for predicting rate of penetration (ROP): A case study from Marun oil field. *Journal of Petroleum Science and Engineering*, vol. 175, pp. 604–623, 2019. DOI: 10.1016/j.petrol.2018.12.013.

[40] M. Anemangely, A. Ramezanzadeh, B. Tokhmechi, A. Molaghab, A. Mohammadian. Drilling rate prediction from petrophysical logs and mud logging data using an optimized multilayer perceptron neural network. *Journal of Geophysics and Engineering*, vol. 15, no. 4, pp. 1146–1159, 2018. DOI: 10.1088/1742-2140/aaac5d.

[41] M. Sabah, M. Talebkeikhah, D. A. Wood, R. Khosravanian, M. Anemangely, A. Younesi. A machine learning approach to predict drilling rate using petrophysical and mud logging data. *Earth Science Informatics*, vol. 12, no. 3, pp. 319–339, 2019. DOI: 10.1007/s12145-019-00381-4.

[42] M. Anemangely, A. Ramezanzadeh, M. M. Behboud. Geomechanical parameter estimation from mechanical specific energy using artificial intelligence. *Journal of Petroleum Science and Engineering*, vol. 175, pp. 407–429, 2019. DOI: 10.1016/j.petrol.2018.12.054.

[43] C. J. Willmott. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, vol. 63, no. 11, pp. 1309–1313, 1982.

Maryam Aljanabi received the B. Eng. degree in computer engineering from Omar Almukhtar University, Libya in 2017, and the M. Sc. degree in computer science from the Applied Science Private University, Jordan in 2020.

Her research interests include machine learning and its applications, artificial intelligence, environmental science, and data science.

E-mail: maryam.aljanabi@asu.edu.jo
ORCID iD: 0000-0003-2152-0788

Mohammad Shkoukani received the B. Sc. degree from Applied Science Private University, Jordan in 2002, and M. Sc. degree from Arab Academy for Banking and Financial Sciences, Jordan in 2004, both in computer. He received the Ph. D. degree in computer information systems from Arab Academy for Banking and Financial Sciences, Jordan in 2009. He is an associate professor at Applied Science Private University, Jordan.

His research interests include agent oriented software engineering, information systems security, and machine learning.

E-mail: m.shkokani@asu.edu.jo (Corresponding author)
ORCID iD: 0000-0002-9401-562X

Mohammad Hijjawi received the Ph. D. degree from Manchester Metropolitan University, UK in 2011. He is an associate professor in Computer Science Department, Faculty of Information Technology, Applied Science Private University, Jordan. He has previous computing based training in several domains. He also acts as the Faculty of Information Technology Dean at Applied Science Private University, Jordan in 2015.

His research interests include natural language processing and machine learning.

E-mail: hijjawi@asu.edu.jo